An Interop Publication

CONNEXIONS®

The Interoperability Report

October 1991 *Special Issue: INTEROP 91 Fall Companion* Volume 5, No. 10

ConneXions — The Interoperability Report tracks current and emerging standards and technologies within the computer and communications industry.

In this issue:

ConneXions is published monthly by Interop, Inc., 480 San Antonio Road, Suite 100, Mountain View, CA 94040, USA. 415-941-3399. Fax: 415-949-1779. Toll-free: 1-800-INTEROP.

Copyright © 1991 by Interop, Inc. Quotation with attribution encouraged.

ConneXions—The Interoperability Report and the *ConneXions* logo are registered trademarks of Interop, Inc.

ISSN 0894–5926

From the Editor

Welcome to INTEROP 91 Fall! This special issue of *ConneXions* is designed to give you background information on some of the tutorials, sessions, and special demonstrations that are taking place during the conference. For those of you who are new to *ConneXions*, this monthly technical journal covers all aspects of computer networking and interoperability.

Our first article is an in-depth look at the many LAN/WAN options facing designers of enterprise networks. Written by Jim Herman and Nick Lippis, the article covers the many new services that are being introduced for building internets such as Frame Relay, SMDS, high-speed circuit-switched data services, and public IP services.

Last month we brought you an article about the *Fiber Distributed Data Interface* (FDDI) demos that are taking place on the exhibit floor. This month, Mark Wolter gives an overview of developments in the FDDI standards arena.

This is followed by a look at SMDS, the *Switched Multimegabit Data Service*. SMDS is another technology which is being demonstrated at the INTEROP 91 Fall exhibition. The article is written by Padma Krishnaswamy and Mehmet Ulema from Bellcore. You can also learn more about SMDS in the conference sessions, check your program guide.

INTEROP® 91
7–11 October 1991
San Jose, California
Convention Center
FALL

Jeff Mogul of Digital Equipment Corporation's Western Research Laboratory will be chairing a session on tools for network monitoring and control. We asked him to summarize some of the important issues in an article. The article appears on page 36.

On Thursday evening, from 5:30pm to 7:30pm, INTEROP will host *The Great IGP Debate* to expose the two prevailing approaches to Interior Routing Protocols, namely IS–IS and OSPF. For background reading we bring you two articles by the proponents of either side.

Privacy Enhanced Mail (PEM) is reaching maturity in the Internet standardizations track. At INTEROP 91 Fall there will be a technology demonstration of PEM. Jim Galvin, of Trusted Information Systems gives a brief overview of the deployment issues for PEM.

That brings us to the end of a not-so-typical issue of *ConneXions*. We have much more in store for you in future issues, and encourage you to subscribe at the special conference discount rate.

Widening Your Internet Horizons:
Wide-Area Options for Internets

by
Nick Lippis, Strategic Networks Consulting, Inc.
and
James Herman, Northeast Consulting Resources, Inc.

The 1980s

Very few network applications are limited to a single isolated site and, increasingly, organizations are turning toward internets to unite diverse user communities across regional, national, and international borders. During the 1980s, a number of different private, wide-area networks (WANs) were popular. These WANs differed in the level at which they operated (transmission, network-layer switching, or a complete proprietary architecture) and the degree to which they were shared by different parts of the organization. Three WANs were typical of a medium to large sized enterprise in the 1980s:

- Single protocol proprietary based wide area nets. An SNA network based on use of relatively slow speed leased, dedicated point-to-point and multidrop circuits connecting cluster controllers and front-end processors, usually operated in the 2.4Kbs to 56Kbs range. Supported only IBM's SNA protocols. Single protocol DECnet networks using primarily point-to-point circuits operating at 56Kbs and below interconnecting DEC computers were also in operation within this time frame.

- An X.25 network based on leased, dedicated point-to-point circuits operating at 9.6Kbs or 56Kbs and shared by many different devices through the use of an X.25 packet switch or *Packet Assembler/Disassembler* (PAD). X.25 provided switched virtual circuit service with very high reliability and could be shared by many different higher level protocols like SNA, DECnet and TCP/IP. [1, 2]

- A T1 network based on leased, dedicated point-to-point circuits operating at 1.5Mbs and shared by many different devices through the use of a time division multiplexer. T1 multiplexers usually broke the T1 down into 64Kbs channels. The most popular use of T1 networks was to cut the cost of leasing many slower speed lines between two locations. Voice and data circuits were usually combined onto a single T1 in order to afford it.

Each of these enterprise WANs is now waning in popularity due to the ascendancy of a new type of WAN, the *enterprise internet.* An internet is a network of networks formed by using routers to interconnect the different component networks. As LANs have become very popular, they have become the primary building block of enterprise internets, although some internets also make use of the older WANs, in particular X.25. As long as two LANs are at the same site, they can be connected merely by plugging a router (or a bridge) into both LANs. If the two LANs are at different sites, however, you now have a WAN problem—how to connect the routers at each site.

The wide-area aspect of internetworking presents network designers with a rich and complex body of issues that include regulatory, political, economic, and technological considerations. On average, 80 percent of an internet's 5-year cost, excluding the cost of labor, goes to wide-area services. Thus, the job of choosing equipment and services for wide-area internetworking should be taken very seriously.

Terminology

Understanding WAN options requires definitions of a few terms and explanation of a couple basic concepts. There are two great dichotomies in wide-area networking:

- *Dedicated versus switched:* Figure 1 shows how the wide-area interconnect would be handled by each type of service.

  A dedicated circuit is one which has a fixed amount of bandwidth capacity (i.e., operates at an unchanging speed) and connects two specific locations. Thus, you need a circuit for each pair of sites that wish to communicate. Dedicated circuits are available in a range of speeds today from 1.2Kbs to 45Mbs.

  A switched network service is one which connects one site to many different locations and may be variable in capacity. Simple dial-up telephone service is the most widely used example, but X.25 and Frame Relay are also switched services. Switched services are traditionally depicted as a cloud. Inside the cloud are switches connected by dedicated circuits.

- *Public versus Private:* A switched network can be implemented as a private network by purchasing the switches and interconnecting them with dedicated circuits. Carriers also provide switched services, usually charging for them based on how heavily they are used. A major issue facing network planners is whether to use public switched services or build a private network to provide those services. Dedicated lines are public, although satellites offer a private implementation for some companies.

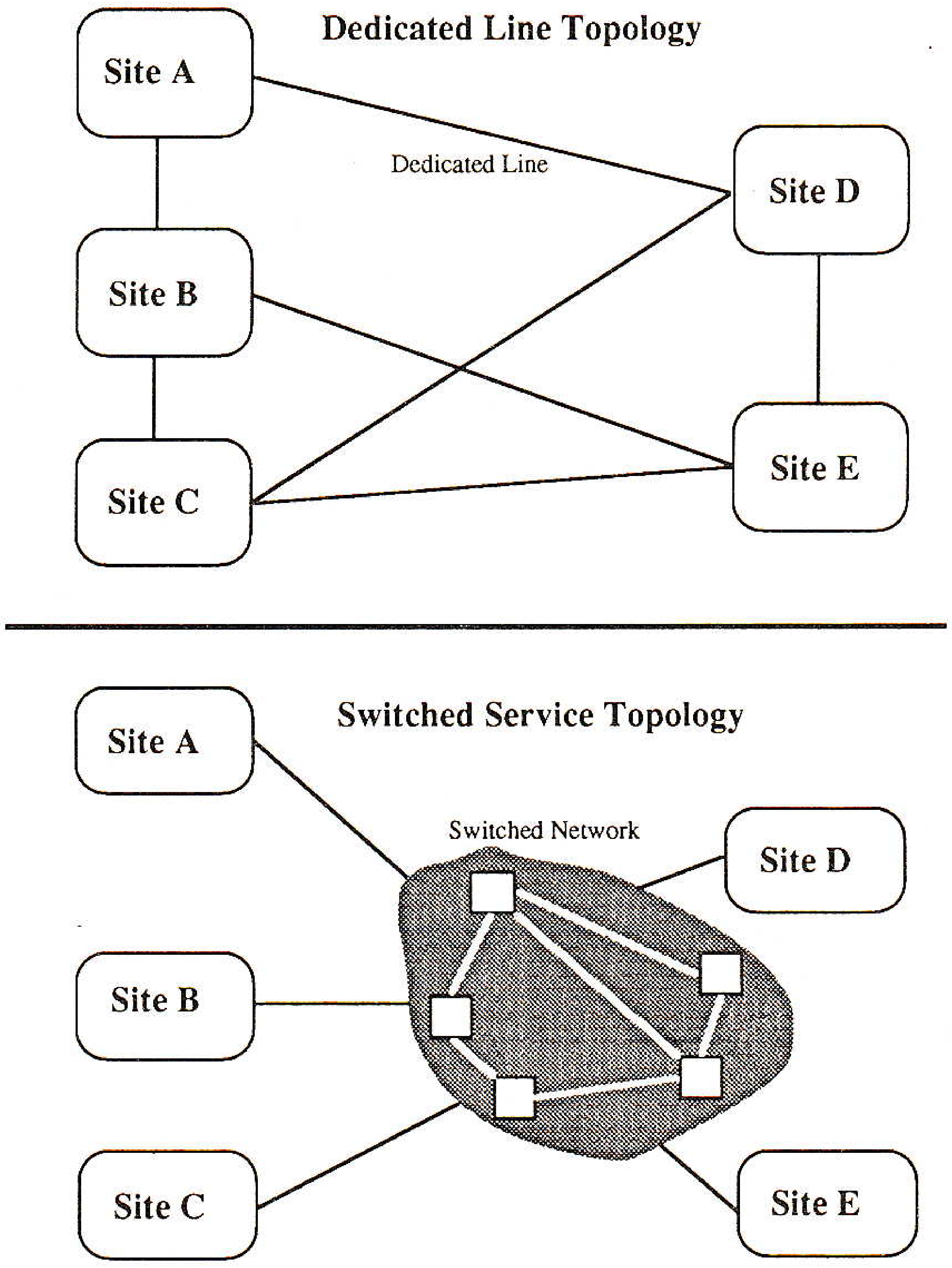

Figure 1: Dedicated versus Switched topologies

Most organizations are going to use public carriers to meet their WAN needs in one way or another. You can just use dedicated circuits leased from the carriers or you can use a carrier switched service. If you wish to build a private switched network, you will still need to lease the circuits between the switches from a carrier, unless you can use private satellite or microwave.

continued on next page

Widening Your Internet Horizons *(continued)*

Public carriers

The public communications carrier industry in the US is extremely complicated due to regulatory issues. Carrier services are divided into *Local Exchange Carriers* (LECs) and *Inter-exchange Carriers* (IXCs), see Figure 2. The majority of LECs are owned by the *Regional Bell Operating Companies* (RBOCs). The three major IXCs are AT&T, MCI and US Sprint. LEC service in the US is regulated so that a LEC can only provide services within the boundaries of a *Local Access and Transport Area* (LATA), which is "roughly" speaking the same as the geographic locale that shares an area code. If you need inter-LATA service, such as New York City to Boston, you must go to an IXC, as well as the LECs in each city, which provides you an access circuit to the IXC. Thus, an inter-city service requires the use of at least three parties, unless you can get to the IXC's *Point Of Presence* (POP) in your LATA directly using microwave or some other kind of bypass technology. Building your own fiber access to a POP is an option, albeit expensive at approximately $13/foot.

LECs and IXCs are the traditional US carriers. A new generation of alternative carriers are coming on the scene now to challenge the monopoly of the LECs. These so-called bypass carriers like Metropolitan Fiber Systems (MFS) and Teleport offer fiber-based services that compete with your local telephone company. They may be leaders in providing dynamic, metropolitan area network (MAN) services. Another kind of carrier is the value-added carrier that provides a switched network service that usually includes some form of protocol conversion, electronic mail or electronic data interchange. CompuServe, US Sprint International (formerly Telenet) and BT Tymnet are leading examples in the US.

In the sections that follow, we will discuss dedicated line services (T1, Fractional-T1, DDS, Hubless DDS, T3) and switched services (X.25, Frame Relay, SMDS, public internet service).

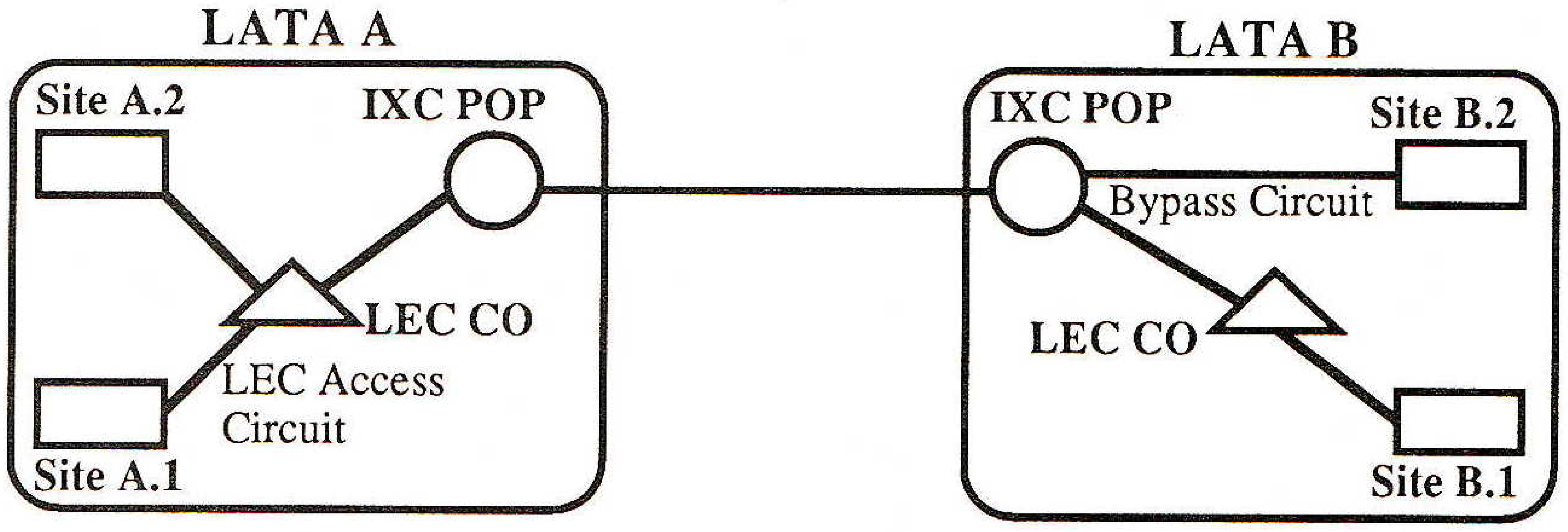

Figure 2: LECs, LATAs and IXCs.

Dedicated lines

If you only have a few sites, with relatively low traffic requirements, you will probably use dedicated lines to link your bridges or routers. Basically, you will go to the appropriate LEC if the sites are within the same LATA, or an IXC if they are in different LATAs, order the circuits you need and then start complaining about how much it all costs.

Traditionally, bridges and routers have been supplied with simple synchronous V.35, and EIA-232-D, RS-422-A, and RS-423-A interfaces. This allowed them to interconnect into the wide-area via a modem or DSU/CSU at dedicated private line rates of 9.6Kbs to T1. Due to the topology updating procedure limitations imposed by spanning tree and source routing, bridging devices do not efficiently support switched services such as dial-up, switched-56, or X.25.

Routers, on the other hand, support a fuller range of wide-area interfaces, including both dedicated lines and switched services. A router's ability to dynamically update its routing tables based on frequent topology changes allows it to handle switched services. In addition, its understanding of network layer addressing enables it to make full use of virtual circuit services such as X.25 and ISDN. Consequently, routers are the center of attention for many of the new wide-area services: Frame Relay and SMDS in particular. (Nearly all router vendors have announced support for Frame Relay, for example.)

Today, the distinction between bridges and routers is disappearing as most new internetworking switches provide the functionality of both, as well as the ability to use high speed services such as T1 and T3 directly, without having to go through a multiplexer. These hybrid devices can be called *Internet Nodal Processors* (INPs) in recognition of their central role in the enterprise backbones of the 1990s.

Point-to-point wide-area services

Shopping for cost-effective WAN services is complicated by the large number of leasing service/options from various carriers, both the IXCs and the LECs. For the small site, with less than 10 people, there are several cost-effective options available today, including Fractional-T1 with analog, hubless *Dataphone Digital Service* (DDS) or DDS access, and in the future, Narrowband ISDN. Narrowband ISDN, depending on its deployment, may provide a promising LAN-WAN interconnect option for small and large internets, particularly if LAN internetworking devices support the LAN-to-ISDN interface.

For connections between large sites (500+ people), T3 (45Mbs) is available in major metropolitan areas throughout most of the United States. A T3 can be leased for as little as the cost of five T1s, depending on mileage. The vendor interest in T3 is considerable and growing. These faster internets are based on INPs interfacing into the telecommunications networks via T1, Fractional-T1, and approaching affordable T3 services.

The confusion in the wide-area data market is heightened by the continued decline of service costs, allowing the wide-area traffic requirements of internets to be met with less expensive WAN bandwidth. For instance, internets running at 56/64Kbs during the mid-1980s will soon be operating at T1. For internets not needing a full T1, Fractional-T1 services provide an excellent transition to higher WAN speeds.

From DDS to T1

Internetworks have traditionally been based on 56Kbs DDS services from the RBOCs and IXCs. Just a few years ago, analysts would have scoffed at the idea of connecting LANs with a full, dedicated T1 circuit. But with the cost of 56Kbs DDS vs. T1 (intra- and inter-LATA) falling dramatically in the period from 1985 to 1990, that has all changed. Figure 3 illustrates the break-even curve between intra-LATA 56Kbs DDS and T1. Intra-LATA communications could represent the cost of a circuit either between locations within a LATA or from a customer location to an IXC POP. From this chronology of cost changes, the following observations can be made:

- Approximately two to five 56Kbs DDS circuits could be displaced by one T1 circuit in the zero-to-100 mile range
- This curve has shifted down from last year by one full 56Kbs circuit, making T1 more economically attractive to high-speed data communications-oriented users
- The biggest decreases over time have been for the shorter distance circuits (0–4 miles) while the 8–100 mile segments have decreased at a more gradual rate

continued on next page

Widening Your Internet Horizons *(continued)*

Therefore, for as little as two 56Kbs DDS circuits (112Kbs) a T1 circuit (1544Kbs or twenty two times the amount of bandwidth) could be purchased. Figure 4 illustrates the inter-LATA break-even curve over time for 56Kbs versus T1. Inter-LATA communications represents the cost of a circuit between LATAs. This is the IXC cost i.e., AT&T, MCI, US Sprint etc. From the curves in the figure it can be deduced that:

- Approximately three to four 56Kbs DDS circuits can be displaced by one T1 circuit within the 25 to 3000 mile range
- This curve has shifted down from last year by nearly two 56Kbs circuits making T1 more economically attractive to high-speed data communications-oriented users
- The largest decrease over time has been for the longer distance-circuits, making cross-country T1 data networks more economically attractive year after year
- The break-even between 56Kbs and T1 is becoming distance-insensitive, and dominated by the access cost

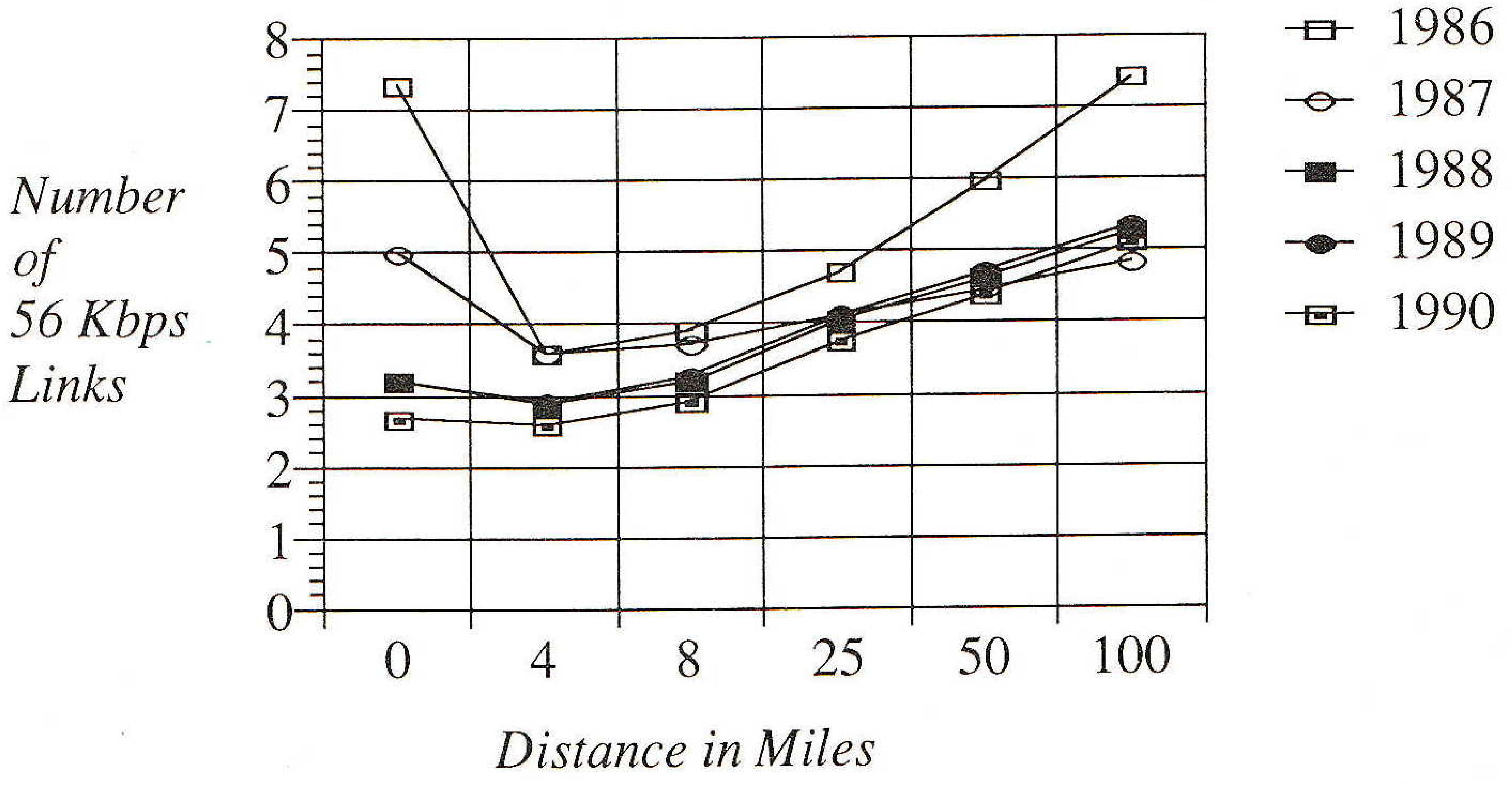

Figure 3: Intra-LATA 56Kbs versus T1 cost over Time

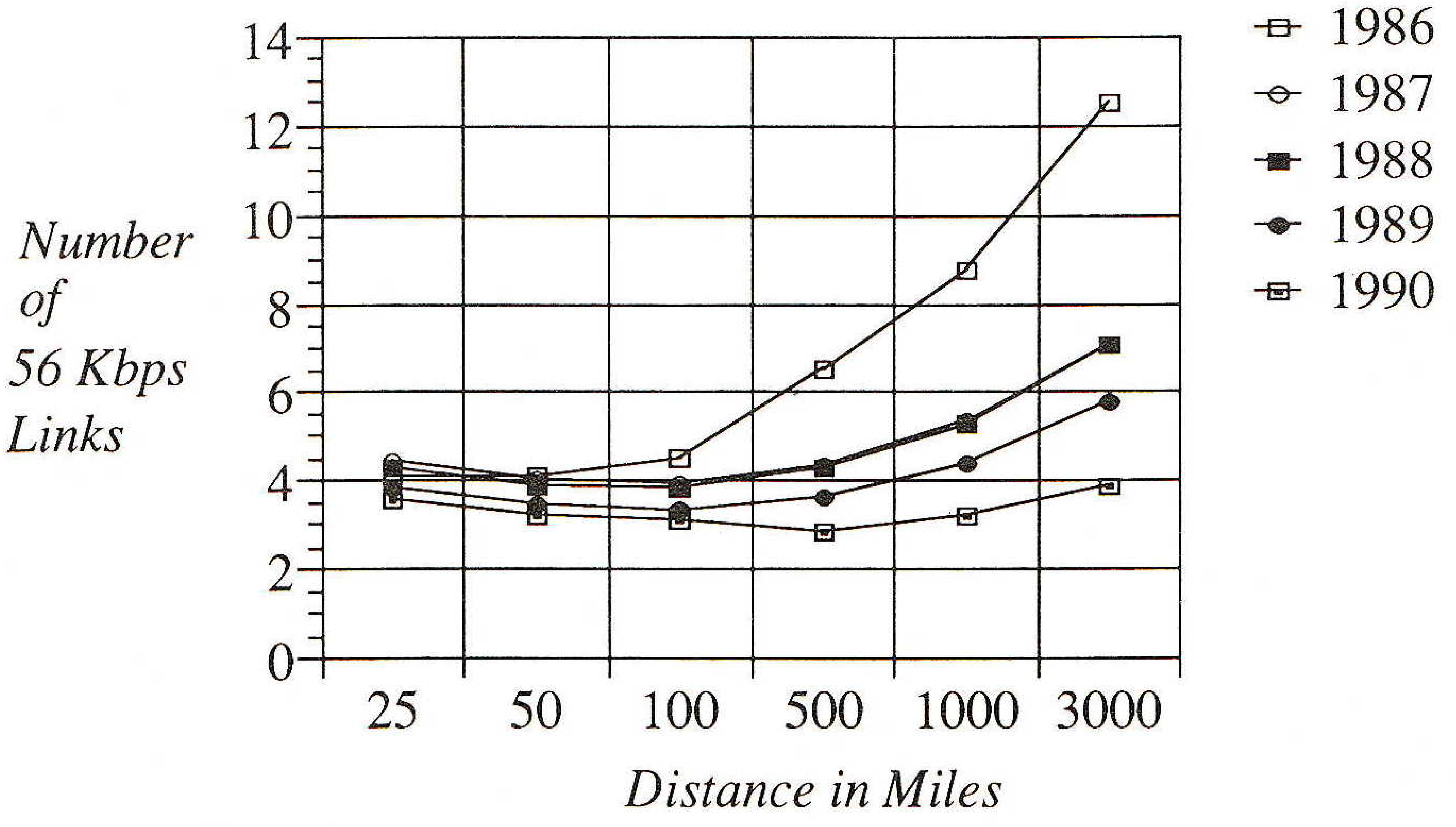

Figure 4: Inter-LATA 56Kbs versus T1 cost over Time

Hubless DDS displaces analog lines

A less-costly 56Kbs service has become available from nearly all of the RBOCs during the past few years. These services are generically called hubless DDS services and offer price savings from regular DDS and analog private lines in the 30 to 60 percent range.

Hubless DDS is offered primarily within a LATA jurisdiction, therefore, being a Local Exchange Carrier or RBOC service. Wherever DDS is offered today, so is hubless DDS. Hubless DDS offers a lower guaranteed quality of service, typically in the 10^{-6} *Bit Error Rate* (BER) range, for a lower price. Alternatively, hubbed DDS offers a higher quality of service, through 125 network monitoring hubs located across the country which are manned 24 hours per day, 7 days per week, providing a BER in the 10^{-9} range.

With the lower cost of hubless DDS services, slower speed analog 2.4Kbs and 9.6Kbs links can sometimes be increased to 56Kbs at the same or slightly higher cost.

Fractional-T1

For those users who don't need a full T1 bandwidth for their data networks and whose circuits are inter-LATA, *Fractional-T1* (FT1) services can offer another level of interconnect flexibility, as well as reduced cost. FT1 offers public multiplexing at the Point of Presence (POP) of the IXC carrier's network. Public multiplexing is based on AT&T's M-24 de facto standard. This multiplexing standard is supported in most *Digital Access Cross-connect Systems* (DACS). DACS is used by nearly all carriers to manage circuits entering and leaving a central office. It is the use of DACS equipment which allows the IXC to offer a variety of access to FT1 services, such as analog private line, 56Kbs, or T1.

Unfortunately, only one LEC has offered an FT1 service today, New England Telephone. The key reason why most LECs do not offer an intra-LATA FT1 service is that the price points between 56Kbs and T1 are too small, i.e., there is not enough margin between the services to make FT1 attractive without cutting into their 56Kbs market.

FT1 service is an inter-LATA service which is priced attractively for a customer profile of geographically distributed sites requiring 64Kbs or higher data transmission. FT1 services—available from AT&T, MCI, US Sprint, Williams Telecommunications Group (WilTel, Tulsa, OK.), and others—offer economic inter-LATA transmission service for smaller sites that need access to the corporate backbone network. FT1 provides a user with *n* times 64Kbs clearchannel circuits with a linear pricing schedule based on the fractional use of T1 bandwidth.

Vendors are announcing FT1 support for their high-speed INP products. For example, Wellfleet Communications Link Node and Concentrator Node routers, Codex's Etherbridge 6310 and CrossComm directly support FT1 today. The FT1 LAN interface is an efficient, single point of attachment into the wide area. A direct FT1 interface into a multiprotocol router combines maximized efficiency of protocol integration over the wide area with a cost-effective transmission option.

The FT1 interface on the INP must be able to multiplex and demultiplex FT1 circuits and map them onto routing addresses. This new interface must also be DACS compatible. Also, there are differences between FT1 carrier offerings which should be understood before making a decision to directly interconnect LANs over FT1 services.

continued on next page

Widening Your Internet Horizons *(continued)*

These differences should be accommodated in the FT1 interface card. For example, some differences between services are:

- Support of either contiguous or non-contiguous channels. For example, a 256Kbs channel may be delivered on timeslots 2, 5, 10, 13 or on 1, 2, 3, and 4
- Support of clear channel services (64 versus 56Kbs channels)
- Differing requirements for filling unused channels
- Bandwidth delivered in channelized format vs. fully transparent higher bandwidth channels

Another feature which could be accommodated in the *Channel Service Unit* (CSU) is support of the *Extended Superframe Format* (ESF). Also, users should investigate vendors' claims to supporting FT1. For example, IBM's midrange *Front End Processor* (FEP) models 3745-170, 3745-150, and 3745-130 are said to support FT1 today, however, an external DSU/CSU or T1 multiplexer is still required. The high-speed scanner does not provide the necessary multiplexing function. (The CSU is the digital network "demarkation point").

Internetworking with T3

While 56 and 64Kbs internetworks are migrating to T1, high-end T1 networks are possible candidates for T3 (44.736Mbs) internetworking. T3 product introductions from companies such as Cisco, StrataCom (the IPX Fastpacket Switch), and potentially IBM are driving this trend. Figure 5 illustrates a potential topology for LAN-to-WAN interconnect at T3 rates.

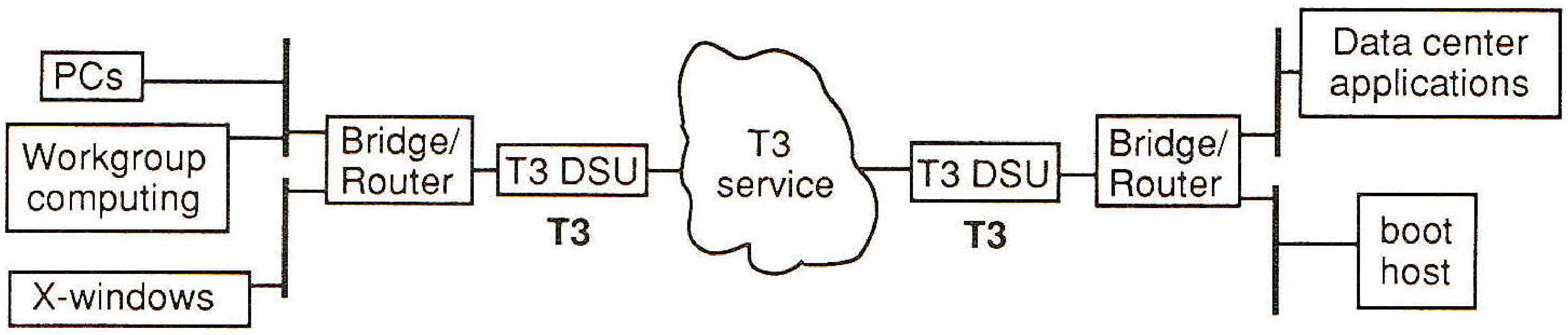

Figure 5: Potential T3 topology

Companies already supporting LAN-to-WAN interconnect at T3 rates include Ultra Network (DS3 Remote Link), Network Systems (FE648 multiprotocol router), Artel Communications Corp., located in Hudson, MA. (MANbridge), and FiberMux Corp., located in Chatsworth, CA. (FX4400). Further, T3plus, Kentrox Industries (Portland, OR), and Digital Link Corp. (Sunnyvale, CA) are providing T3 DSU/CSUs which facilitate interconnection into the wide area by providing a modified V.35 interface to the internetworking devices. This V.35 interface, called the *High Speed Serial Interface,* (HSSI) is being standardized by DSU/CSU vendors in cooperation with a few internetworking vendors. HSSI has also shown up in the new Adaptive SONET Transmission Manager (STM). In addition to its T3-capable router, Network Systems is currently selling a "proprietary" T3 DSU/CSU.

Four categories

There are essentially four categories of product development which define T3 networking alternatives. First, there are the internetworking vendors extending their existing INP platforms into the high-speed T3 environment to meet FDDI-to-T3 and LAN-to-T3 requirements.These vendors are Network Systems, Ultra Network, and Cisco. While IBM does not have a T3 internetworking device today, they have demonstrated a T3 interface on the RISC System (RS) 6000 which provided FDDI-to-T3 routing with potential application for the NSFNET. [3]

In the second category are the proprietary T3 LAN-to-WAN interconnect vendors such as Artel and FiberMux. The internal switching fabric, as well as LAN access into the WAN, is proprietary for these vendors, not the T3 interface itself. For example, Artel provides a proprietary appended token ring as a wide-area architecture while FiberMux provides a proprietary *Time Division Multiplexing* (TDM) architecture over T3 facilities.

Both Artel and FiberMux provide 802.3 bridging over wide-area T3 links while FiberMux supports Token Ring, Arcnet, T1 voice, IBM channel extension, and a wide variety of other interfaces. The choice of one architecture over another strongly depends on the particular requirement, however. It is very unlikely that these products will find wide acceptance in the emerging T3 market, given the growing choice of standards-based T3 interconnect devices available.

The third category consists of T1 multiplexer vendors such as Infotron, Timeplex, Network Equipment Technologies (NET), and Newbridge Networks. While their products support T3 as well as multiple T3 rates, they do not provide an aggregate 44Mbs between the internetworking device (e.g., router) and the wide area. The aggregate data rates available to a single device are generally 2Mbs due to the bus bandwidth of these devices.

A fourth category is represented by a newer class of devices, including Adaptive's SONET Transmission Manager (STM), T3plus's T3 Bandwidth Manager (BMX45) and Digital Link's DL3000 T3 mux. These devices provide T3 bandwidth up to the full 44Mbs in user-specified allocations controlled through a management console. This allows connection management for chunks of high bandwidth between various devices for voice, data, and video. These devices are to T3 what the first point-to-point T1 multiplexors in the mid 1980s were to T1, in terms of providing voice and data integration and limited networking functionality.

It is not quite fair to equate the DL3000 and BMX45 with the Adaptive STM, considering that the STM handles more lines and provides a great deal more functionality—such as support for the Q.931 protocols—which allows it to participate in the new switched T1 and T3 services.

Vendor	*Product*	*Type*	*Framing Protocols*	*Application*	*Throughput*
NET	IDNX 90	Mux	Proprietary	Integ. Voice+Data	~2Mbs
Timeplex	TX3	Mux	M13	Integ. Voice+Data	~2Mbs
Newbridge	Mainstreet 3645	Mux	M13 or Prop.	Integ. Voice+Data	~2Mbs
Infotron	Streamline 45	Mux	M13 or SYNTRAN	Integ. Voice+Data	~2Mbs
ADAPTIVE	STM	SONET Mux	SONET	Integ. Voice+Data	44Mbs
T3plus	BMX45	Mux	Proprietary	Integ. Voice+Data	44Mbs
Cisco	AGS+	INP	HSSI	LAN Interconnect	44Mbs
Ultranet	DS3 Remote Link	Router	HSSI	LAN Interconnect	44Mbs
Network Systems	FE 648	Router	Proprietary	LAN Interconnect	44Mbs
Artel	MANbridge	Bridge	Proprietary	LAN Interconnect	44Mbs
FiberMux	FX4400	Mux	Proprietary	LAN Interconnect	44Mbs

Table 1: T3 Products

Table 1 lists vendors by the four major categories of T3 functionality. Within this table only categories 1 and 4 will address the growing LAN interconnect requirements of the 1990s.

continued on next page

Widening Your Internet Horizons *(continued)*

Category 1 provides attachment from various LANs into a high speed serial interface formatted for the telecommunications network. Category four allows for the multiplexing of various category-one devices for ultra-high bandwidth requirements.

T3 service is available in major metropolitan areas throughout most of the United States. The price of T3 is surprisingly low and can only continue to fall with competition between the RBOCs and alternative access carriers driving prices down. The total price of T3 depends on the access method employed. The break-even between inter-LATA T1 and T3 services taking into account various access arrangements are:

- Approximately five to fourteen T1 circuits can be displaced by one T3 circuit within the 25-to-3000 mile range, when access is provided through a T3 access tariff from an LEC, for example.
- Customer-provided fiber optic access to an IXC POP dominates the total cost of T3 service. For customer-provided access, approximately fourteen to sixteen T1 circuits can be displaced by one T3 circuit within the 25-to-3000 mile range—i.e., the curve is almost mileage-insensitive.
- Customer-provided T3 microwave access would produce a break-even curve in between the fiber optic and tariffed provided access curves.
- These two curves have shifted up from last year by four to six T1 circuits. This is due to the decreasing cost of T1 rather than increasing T3 cost.

The last observation is an important one. The cost of T3 tariffs have not fallen as fast as T1, therefore it may in the short run be more difficult to cost justify T3 services. Alternative access providers may, however, drive down T3 access cost through competition if their regulatory efforts (discussed below) prove fruitful.

Fractional-T3

Fractional-T3 (FT3) is not a particularly compelling LAN interconnect option, in spite of the significance of its little brother—FT1. There are two companies offering FT3 services today, WilTel and Cable and Wireless Communications Inc. (C&W, Vienna, VA.). AT&T and MCI, as well as LiTel Telecommunications Corp. (Worthington, OH) have voiced interest in offering, but have made no firm commitments. FT3 is much like FT1, however, it offers fractions of T3 bandwidth at the T1 increments up to 28 circuits, tariffed on a linear pricing schedule depending on the number of T1s and distance.

In effect, FT3 service at this point is nothing more than a discounted pricing schedule for bulk T1 lines. One of the major drawbacks of today's offerings from WilTel and C&W is that the T3 circuit is channelized, meaning that the only way to access bandwidth is in chunks of T1. Consequently, one cannot allocate 10Mbs for an 802.3 INP or 4Mbs for a 802.4 INP, only channels of T1.

The next generation of FT3 services may prove to be more exciting, allowing the partitioning of T3 bandwidth in unchannelized allocations, as well as providing the user with control over bandwidth provisioning. What is delaying the deployment of this type of service is the lack of feature-rich carrier transmission equipment.

SONET

SONET (*Synchronous Optical NETwork*) is an emerging hierarchical high speed multiplexing international standard which will allow the carriers to effectively exploit the huge installed base of fiber optic transmission.

Most importantly, SONET will break the 45Mbs speed barrier, which is the largest bandwidth the carriers can offer today as a private line service.

With the introduction of SONET carrier equipment in the 1992 time frame, the carriers will be able to offer an entirely new range of high speed public private line and multiplexing services. The OC-3 or *Optical Connection 3* level of the SONET hierarchy, operating at approximately 155Mbs, has attracted the product development efforts of many internetworking, telecommunications, and telephony manufactures, therefore, leading us to believe that OC-3 will most likely be the preferred high speed premises-to-carrier interface in the early and mid 1990s. Therefore, it is SONET and not T3 which will ultimately offer high speed LAN interconnect services for the 1990s.

Switched WAN services

A variety of switched services are also available, again in either a public or private implementation. The switched services of the past (X.25 and dial-up) are generally too slow to meet the needs of today's internets. Now, a new generation of switched services is coming to market. These new services have been expressly designed for the LAN interconnect problem and offer high speed, bandwidth on demand and, presumably, affordability.

Public switched services offer a significant reduction in complexity for the enterprise network manager since they put the problem of leasing most of the circuits in the hands of the carrier. Using public switched services in the past, however, has also meant a drop in service quality since the shared service cannot be tuned to the exact needs of any one customer's applications. Also, switched services have tended to be priced on usage. This means that they are generally favorable economically for low volume use by small offices rather than as a backbone option. The pricing on the new switched services is not settled yet and it remains to be seen whether they will become viable backbone options.

X.25: A limited role

Although used successfully for terminal networks and other low-speed applications, public X.25 networks are too slow and overhead-laden to become a mainstream WAN alternative for internets. For some smaller LAN interconnect uses, X.25 is a good fit in the near term. But private and public X.25 networks will not find much acceptance for internet architectures as the '90s proceed, due to the introduction of Frame Relay-based private and public network equipment and services (discussed in the next section) International applications of X.25 will still be important, however.

In a low-volume situation with many scattered sites, a public X.25 packet switch network may be the most appropriate and economical. In such a case, public X.25 packet switch vendors such as Sprint International (Reston, VA) and BT Tymnet Inc. (San Jose, CA) rent the use of their switches to customers on a per-use basis.

Public X.25 networks have never been extremely cost-effective for speeds above 56Kbs. Public X.25 nets operating at 56Kbs access are more expensive even than 56Kbs private-line configurations, therefore, providing very little incentive to use this service. As a result, many corporations will opt for the use of public X.25 networks to integrate the small site into the corporate backbone and for international connectivity leaving private X.25 networks as an anomaly.

continued on next page

Widening Your Internet Horizons *(continued)*

Other reasons why private X.25 networks will not find wide acceptance as support for the internet architecture:

- X.25 was developed with the problems of analog transmission in mind; the overhead was designed to compensate for this. This is now moot in the US since transmission telecommunication systems are nearly all digital (excepting the local loop). This is precisely why X.25 is still important overseas and internationally.
- With the improvements in both telecommunication transmission systems and end system protocols, the built in overhead of X.25's sophisticated error correction and congestion control mechanisms providing a highly reliable service is not needed any longer for domestic US applications.
- X.25 mainly operates at below 56/64Kbs, therefore limiting its scalability in large corporate internets.

Frame Relay

Recently, a new technology for dynamically allocating bandwidth to bursty internet traffic—*Frame Relay*—began emerging at a lightning pace, promising LAN performance on the WAN.

Frame Relay is one of the most-discussed internet interfaces (and public services) in years. It promises to provide improved wide-area network performance, streamlined interconnect, efficient use of wide-area bandwidth, and reduced overall cost of building and maintaining internets. Numerous vendors from all corners of the industry have announced plans to offer a Frame Relay interface for use in either public or private networks.

Frame Relay has been introduced or announced by all of the major internetworking vendors and many of the packet-switching, T1 mux, central office equipment, and big system vendors. By the end of 1991, Cisco, ACC, Wellfleet, DEC, Network Systems/Vitalink, Proteon, StrataCom, Newbridge Networks, Sprint International, Telematics International (Fort Lauderdale, Fla.), Netrix Corp. (Herndon, VA), Hughes Network Systems (Germantown, MD), AT&T, and Northern Telecom (NTI, Nashville, TN) will have shipped their first Frame Relay interfaces.

As of this writing there are more than 36 vendors who have announced their support for the *Local Management Interface* (LMI) extensions to the Frame Relay interface, spearheaded by the NTI, DEC, StrataCom, and Cisco consortium. IBM has been a strong advocate for Frame Relay carriage of SNA traffic within the ANSI T1/S1 standards committee and has announced support for Frame Relay on its cluster controllers, FEPs, and LAN interconnect bridge.

Frame Relay provides a packet-multiplexed interface between internetworking devices and the WAN, where up to 16,384 (LAPD) *Permanent Virtual Circuits* (PVCs) can be routed across the wide area. Consequently, a single V.35-type interface from an INP into the WAN delivers the 16,384 PVCs plus a control channel to the WAN. The WAN provides dynamic routing of these PVCs between INPs, thus creating a logical mesh topology between packet-forwarding equipment. While the number of PVCs is large, the limiting factor in designing internets with Frame Relay technology will invariably be the access bandwidth.

Conformance issues

Frame Relay has been ratified within ANSI T1S1/88-2242, "Frame Relay Bearer Service—Architectural Framework and Service Description." It is also an emerging standard described by CCITT under its I-Series recommendations as I.122, "Framework for Additional Packet Mode Bearer Services," which may be ratified during 1991. Most vendors are supporting the emerging CCITT and/or ANSI Frame Relay standards, as well as the LMI extensions mentioned above.

The NTI, DEC, Cisco, and StrataCom consortium was formed to reduce the risk of the various versions of the Frame Relay interface on the market not being interoperable. It is unknown as of this writing whether their efforts will be successful. However, all vendors will be supporting the bare minimum of the ANSI and CCITT standards, while most will incorporate the enhanced features detailed within the "Frame Relay Specification with Extensions" document, produced by the consortium members. [4]

A good example of such an enhancement is Cisco's Frame Relay interface (developed with StrataCom), which will support a dynamic address resolution protocol which dynamically maps LAN internetworking addresses to Frame Relay LAPD *Data Link Connection Identifier* (DLCI) addresses. This will enable Cisco routers to use the *Address Resolution Protocol* (ARP) across a Frame Relay interface to map network addresses to physical (Frame Relay) addresses.

Each internetworking device connected into the WAN via the Frame Relay interface appears to have a physical link to each other, and is thus capable of being part of a fully meshed networking topology. Therefore, every internetworking device is logically adjacent in a Frame Relay network. This provides network managers with improved performance, since the processing overhead associated with packets traversing intermediate hops between packet forwarding devices is nearly eliminated.

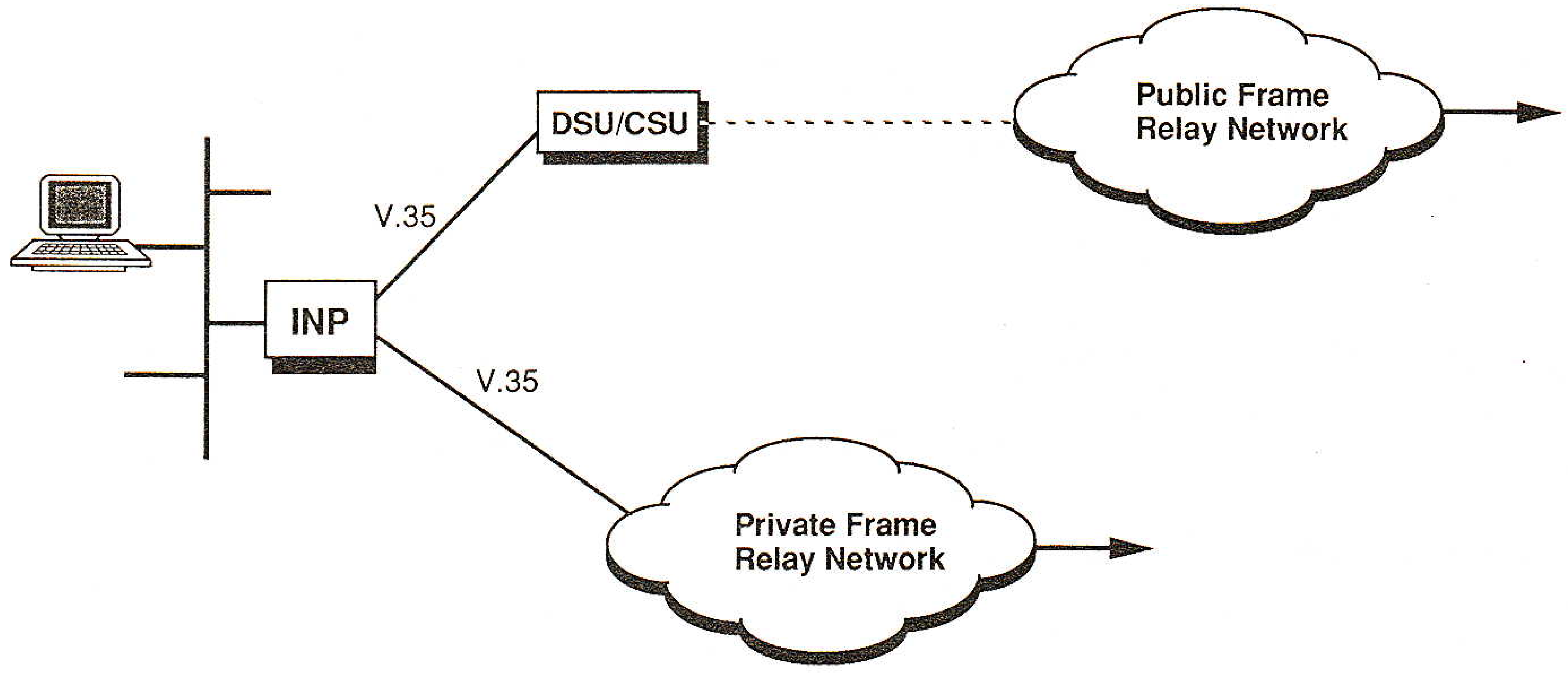

Figure 6: Public and private Frame Relay network topologies.

Private versus Public Frame Relay Networks

As with X.25 networks, Frame Relay is manifesting itself as both a private and public network solution. The standards and conformance work promise improved performance associated with Frame Relay networks, but there are still questions about the differences between private and public Frame Relay networks. US Sprint has announced plans to offer a public Frame Relay network to be available in Q3/91. CompuServe Inc. (Columbus, OH), BT Tymnet, and WilTel have also announced public Frame Relay services to be widely available this year. WilTel's WilPac Frame Relay service is available nationwide and is based on a flat rate tariff structure with various quality of service options.

continued on next page

Widening Your Internet Horizons *(continued)*

The tariff structures for US Sprint, CompuServe, and BT Tymnet Frame Relay services are unknown at this time; however, users should expect that public Frame Relay pricing will be at least 15 percent less expensive than comparable 56 Kbs and T1 networks—which will price X.25 out of the market. Note that BT Tymnet has delayed its public Frame Relay offering for six months.

Figure 6, on the previous page, illustrates both the public and private Frame Relay network topologies. Some of the key architectural benefits of either type of Frame Relay network are:

- Single physical interface into the WAN, which allows a mesh topology
- Scalable to support high speeds (T1+)
- Scalable to support small (3 nodes) to large (1000+ nodes) networks
- Reduced overall cost of ownership (discussed below)
- User networks can be built without detailed knowledge of their traffic requirements
- Improves performance by reducing overall network delay
- Expands and contracts bandwidth based on user demand—particularly well-suited for an internet

There are drawbacks to the emerging Frame Relay networks as well. The *Switched Virtual Circuit* (SVC) feature will not be available until late 1992 and this will largely limit Frame Relay networks to intra-enterprise use. Also on the downside is their lack of flow and congestion control, since all four flow-control mechanisms in X.25 are eliminated, requiring the devices using the Frame Relay service to limit traffic flow to sustainable levels. This could create compatibility problems between different types of INP equipment using different methodologies to accomplish flow control. There is also a clear lack of coupling congestion information from a Frame Relay network to the upper layer protocols of IP and DECnet by the internetworking vendors. There are three different syntactic methods for congestion avoidance signaling within the Frame Relay standard, but these are all optional for both user and network. This results in considerable opportunity for incompatibility.

Private Frame Relay applications

In the private network scenario, the user owns and controls all of the switching equipment (T1 mux, packet switch, etc.) which provides wide-area services in support of internet traffic. There are six key vendors offering private Frame Relay switching equipment: StrataCom, Hughes, Telematics, Netrix, Newbridge Networks and Network Equipment Technologies, Inc.

Wellfleet has announced that its Frame Relay interface will be free of charge, while Cisco is providing a free upgrade from its X.25 interface to Frame Relay. If a Cisco customer is not currently using its X.25 card, a $750 software upgrade will turn a V.35, RS-449, or RS-232 card into a Frame Relay interface. On the T1 side, StrataCom has delivered its Frame Relay interface, which can be obtained by a $1,500 software upgrade. Thus, the additional capital cost of obtaining Frame Relay is nominal.

The key benefit of Frame Relay in the INP/mux environment is the reduction in the number of interfaces between INPs and the WAN. Note that this property is true for SMDS as well. As illustrated above this will reduce the cost of building mesh internets. This is significant, since most T1 interfaces for routers are priced in the $2,000 range. With one interface to the mux, the INP gets access to a large number of virtual circuits, whereas many expensive physical interfaces between the router and the mux were previously required.

Public Frame Relay applications

The time for planning Frame Relay applications is now, considering that WilTel is currently offering it, and Sprint service will be available in production mode in Q3/91. BT Tymnet's and CompuServe's service will be available in July 1991 and January 1992, respectively. MCI has also announced that it will offer Frame Relay services and AT&T is rumored to be planning its introduction by the end of 1991.

From the user perspective, traditional private line based internet architectures suffer from low average utilization of link speed, as well as costliness and rigidity. FT1 is an exception. A public Frame Relay service promises to reduce ongoing access cost and provide a 15 percent reduction in IXC cost. Also, increased flexibility (in terms of connectivity) is realized through the public network's provision of a logical mesh topology and the ability to mix and match access services (e.g., 56Kbs to T1). The most important increase in performance is the WAN's provision of bandwidth on demand, to fit the bursty nature of LAN interconnect traffic.

Public Frame Relay services may offer a better way of communicating between enterprises when traffic patterns can be somewhat quantified—i.e., between customers, vendors, and *Value Added Networks* (VANs) such as EDI, X.400 etc. The trade-off between private and public Frame Relay networks comes down to a significant reduction in the monthly recurring and capital cost of wide-area bandwidth (no need for T1 nodal processor, i.e., redundant switching), for a partial loss of network control.

The following are some of the potential benefits of a public Frame Relay network. (These properties are largely associated with SMDS as well.)

- Substantial reduction of monthly recurring cost through lower access cost
- Leveraging of the public network's robustness and intelligence
- Economical interconnection of small sites into the corporate enterprise network
- Improved access to customers, vendors, and VANS
- Reduced capital cost
- Reduced complexity (reduced redundant switching)

Before a rational decision can be made when comparing public and private Frame Relay networks, users need to understand the following about public services:

- What management interfaces are available, and what information is provided by the carrier
- What the cost of the service is
- Whether the service is compatible with a given INP Frame Relay interface

continued on next page

Widening Your Internet Horizons *(continued)*

Such questions will only be answered as Sprint International, MCI, AT&T, BT Tymnet, CompuServe, and others actually introduce their Frame Relay services. Sprint has announced that its Frame Relay service will support the LMI extensions, therefore assuring compatibility with the installed base of internetworking equipment. WilTel, as of this writing and although its service is available, has not publicly stated the support of the LMI standard. The potential improvements in cost/performance for users building internets (and the market opportunities for carriers) are so enormous that it is difficult to see how this service could fail to be successful.

Public Virtual Data Networks

Clearly, data is going to be the arena where strategic network decisions are made and where carriers gain or lose control of their accounts. In order for the carriers to succeed on the strategic data communications front, they will have to offer much more value than the static wide-area bandwidth found in traditional T1 and DDS services. To date, the majority of users have purchased their private line data services from AT&T's long-haul division, which controls over 90 percent of the long-haul data market. This will change rapidly.

Dynamic, on-demand bandwidth, geared toward the internet paradigm, is the essence of the new carrier offerings and will provide LAN-like performance in the WAN. One of the key manifestations of this trend is the carrier-provided "virtual data network." The goal of the carriers in this area is to offer data services that allocate bandwidth within milliseconds of demand, thus making the WAN behave more and more like a LAN.

Just as virtual voice networks are bringing voice back to the public network, private data networks will become increasingly hybrid private/public configurations as the 1990s progress. The IXCs and LECs are currently investigating a wide range of technologies that will be the basis of virtual data networks aimed at bringing data into the public network.

The new public data services promise to reduce both access and IXC costs. Access costs represent the largest part of the total cost of WAN backbone bandwidth. Reduction of access charges are realized when internets interface to the public WAN through a "single" access link. In this scenario, the public network provides routing, addressing, and management. The public network hence provides a logical mesh topology and the ability to mix and match access services (e.g., 56Kbs to T1 to T3).

Vastly increased performance for internets is projected as virtual data networks provide flexible bandwidth that can expand and contract in response to the bursty nature of LAN interconnect traffic. The potential benefits of a public data network are many, including:

- Extension of the enterprise network down to the small site
- Economic integration of the small site into the corporate backbone
- Reduction of major capital costs
- Reduction of personnel needs due to partial outsourcing
- Better access to customers, vendors, and VANs as a result of the ubiquitous aspect of public data networks

In spite of their allure, these potential benefits are greeted with skepticism by many network owners. Questions abound: Can the carriers, especially the LECs, sell data services? Can a carrier be trusted with a corporation's competitive advantage? Will services be tariffed attractively? What management interfaces and management information will be provided by carriers? And while public networks offer potential reductions in the monthly recurring and capital costs, the trade-off for the network owner is a partial loss of network control. While this last point hasn't bothered uses on the voice side, the strategic aspect of data networking requires control.

The IXCs, and potentially the RBOCs themselves, are looking toward data services as a major portion of their business. A number of developments are prerequisites for the realization of this shift to public data networks. Above all, the RBOCs must rapidly deploy MANs in major metropolitan areas. MANs make the public network look like a large LAN and hence are perfect for extending the client-server computing environment that is developing on LANs. MAN technology with its high-speed, flexible bandwidth—not ISDN—is the key to a sophisticated role in business data networking for the RBOCs in the 1990s.

Carrier activity in enhanced wide-area technologies is tremendous. If the carriers succeed, internets of the future will be serviced by Frame Relay, IEEE 802.6 data links, SONET, Asynchronous Transfer Mode (ATM), and high-speed circuit switching. These developing technologies are the underpinning for a diversity of deliverable services.

Taken together, these technologies and services are the raw material from which the virtual data networks of the 90s will be crafted. Of immediate importance in the virtual data network market are Frame Relay, MANs, and high-speed switched services. If the carriers can implement viable MANs (and interconnect them across the wide area), data will tend to return to the public network, just as voice has. In the same way carriers use virtual networks to precipitate the dismantling of private voice networks, the virtual network concept can be extended to data.

SMDS

Switched Multimegabit Data Service (SMDS) is a MAN for data communications. Its goal is to extend the low-delay and high-bandwidth performance of LANs to the metropolitan area. To do this the SMDS specification is based on a connectionless datagram service extending 4 and 16Mbs Token Ring and 10Mbs Ethernet LANs into the MAN. SMDS also supports 16, 25, and 34Mbs speeds through fiber optic local access as well as slower speed T1. [5]

Since it is connectionless and high-speed, SMDS is much like a LAN in the MAN. It will be a critical service of the mid-1990s, and thus the RBOCs are seriously interested in its development and deployment. The LECs should be the drivers deploying SMDS in late 1992, whereas the IXCs are the main drivers positioning the deployment of a public Frame Relay service in 1991.

The service concept of SMDS is indeed attractive. AT&T's implementation of SMDS is built around its Datakit products which is currently being re-engineered and will emerge as a private/public broadband networking platform based on cell-relay technology. Siemens Information Systems Inc. (Boca Raton, FL) and Alcatel PABX Systems Corp. (Alexandria, VA) are two other telephony manufacturers which are providing SMDS equipment to the LECs.

continued on next page

Widening Your Internet Horizons *(continued)*

On the customer side, only a handful of vendors—Sun Microsystems, Ungermann-Bass Inc. (Santa Clara, Calif.), Proteon, Cisco, and AT&T —are supporting the IEEE 802.6 connectionless MAC, which may be interfaced into an eventual SMDS service. Note that these vendors have not announced such products as available for commercial use, but did announce and even demonstrate them as proof of concept at the INTEROP 90 trade show.

Some LECs are in a quandary over deployment, since both Frame relay and SMDS support dynamic bandwidth allocation, bursty LAN interconnect, and high-speed point-to-multipoint LAN interconnect over a single local access loop. There is an increasing overlap in Frame Relay and SMDS access speeds, with Frame Relay being extended to 45Mbs and SMDS descending down to T1. The RBOCs seem to have been guided into the right direction by most of their decisions to support both Frame Relay and SMDS. As of this writing, over five of the seven RBOCs have field trials planned for Frame Relay.

SMDS will first be offered at the T1 (1.54Mbs) and T3 (45Mbs) rates through various classes of service (4, 10, 16, 25, and 34Mbs), while a public Frame Relay service will initially be available at 64Kbs to T1 rates. NTI will be demonstrating Frame Relay operating at 45Mbs at the INTEROP 91 conference, thus showing that Frame Relay can be scaled up to 45Mbs, but there are still unanswered technical and implementation questions. These include uncertainties about Frame Relay performance above 10Mbs, due to bit stuffing and HDLC synchronization.

Implementation issues

While at first glance the higher speeds of SMDS (34Mbs being the maximum) may be attractive, the real market for LAN interconnect access is at 64Kbs to T1 rates today. Today, over 25 percent of the installed base of remote bridges and routers use T1 links to interconnect LANs. Another 50 percent use 56/64Kbs services, while nearly 20 percent use Fractional-T1—with the remainder still in the analog stone age. On the other hand, the deployment of FDDI will certainly drive the need for SMDS.

This is not to say that higher-bandwidth LAN interconnect speeds are not in demand. Indeed, in the medical and discrete manufacturing industries, 45Mbs is required. Some of this need, however, can be met by traditional point-to-point 45Mbs services interconnected by a variety of routers which offer T3 interfaces. This was demonstrated at INTEROP 90 by Cisco's AGS+ router interfaced into T3plus's DSU45 DSU/CSU via the new *High Speed Serial Interface* (HSSI) developed by T3plus, Cisco and others. Such a configuration is particularly important for extending FDDI networks into the wide area. It is SMDS's single interface into the MAN which will be a major competitor to point-to-point T3 LAN interconnect (just like Frame Relay's single interface into the WAN is its major advantage over point-to-point T1 today).

Hybrid wide-area internets

The advent of high-speed public data services (Frame Relay, SMDS, and high-speed circuit switched services) means the carrier has a set of tools with which to migrate private data networks into the public network.

With the introduction of public data services from carriers and others, a shift from entirely private networks toward part-public and part-private hybrid data networks will emerge. The growing acceptance of the internet architecture among commercial businesses (retail, insurance, banking, etc.) is a primary impetus toward the development of hybrid networks and more robust data services.

The advent of viable public data service means that network owners will be faced with a new economic model to understand, and possibly exploit. If the incentive is economics, users will have to categorize their data services into strategic and non-strategic corporate resources. A non-strategic service, for example, might be interconnection into the national research Internet, while strategic services might be internal electronic mail, notes, an airline reservation system, etc. Some critical applications will be subject to considerable compromise when given to a carrier to manage. The key trade-off may very well be increased cost savings vs. control. The nature and quality of the network management information provided by the carrier will certainly affect the user's control.

Public internet services

Other potential competitors in this area are public IP services from companies such as Advanced Network and Services, Inc. (ANS), PSINet, and UUNET, for example. ANS, which operates the NSFNET backbone, has recently formed a subsidiary CO+RE to offer public, commercial internet service. ANS and others could relegate the LECs and RBOCs to offering low-margin special-access private lines from customer sites to nodes on their respective networks. These public IP based networks are attracting venture capital and are well positioned to seize the business data market away from the slow moving LECs. [6]

A public internet service recreates the capabilities of the DoD's Internet but without the restrictions on usage imposed by government, but with a for-profit charging policy. Public internets are similar to the X.25 VANs of the 1980s with one big difference. Most of the large vendors did not embrace X.25 very aggressively in the 1980s and encouraged their customers to build private networks. Today, most vendors are actively supporting internetworking and the potential market for the new IP services is huge.

As with the other public services we have examined, the advantage here is that the public IP service provides a single access point through which to reach many sites. The greatest use of the public IP services will most likely be for inter-enterprise internetworking. More and more companies need to share information among each other and since they are all building internets, the need for an inter-enterprise internet service is clear. The first public IP services only handle TCP/IP. In the future, they will most likely expand the range of protocols they handle.

Narrowband ISDN

During the period between late 1992 and early 1996, ISDN *Basic Rate Interface* (BRI) services will become widely available from U.S. RBOCs (approximately 12 percent of the more than 26,000 central offices). In another instance, it is estimated by Northern Business Information Systems that approximately 22 percent of the six million Centrex lines will be ISDN BRI by the end of 1991. [7]

ISDN as a service to interconnect residential and small business sites to corporate and public information services may prove to be very attractive. The ISDN BRI user-to-network interface will first impact dial-up and analog private-line internet users.

continued on next page

Widening Your Internet Horizons *(continued)*

ISDN offers multiple switched services over a single access circuit. It can offer digital dial-up, X.25, and Frame Relay all over the same circuit, if needed. ISDN will eventually be available almost everywhere in the US, since the LECs are deploying it fairly universally. It will still be a number of years, however, before it could be called ubiquitous.

Summary

As internet architectures take off, there is an enormous opportunity for public carriers to offer diverse local and national data services. There are many new services being introduced that both increase the bandwidth available for building internets and potentially lower the cost. Private enterprises must once again reevaluate the mixture of private and public networking that makes sense for them as Frame Relay, SMDS, High Speed Circuit Switched Data Services, and public IP services become widely available.

Network speeds will increase dramatically over the next few years as the Internet becomes the primary enterprise-wide network. Today's 56Kbs circuits will be upgraded to T1 and Fractional-T1. Today's T1 will be upgraded to T3 and then SONET OC-3. For those small sites, hubless DDS at 56Kbs or Basic Rate ISDN will be the most cost effective options.

In all of this, it is important to remember the need for network management of the WAN. Carriers are offering increased management information about their services. Linking carrier management information into private management systems will be a major initiative in the next few years.

In the end, a hybrid public-private internet architecture is the likely outcome of these developments. Wherever there is a very high concentration of traffic, private solutions will always be better, but for the majority of sites a public service will be more cost effective and certainly easier to manage. If the carriers can win over the confidence of users, there is a great new opportunity in internetworking awaiting them.

References

[1] D. Vair, "Components of OSI: X.25—the Network, Data Link, and Physical Layers of the OSI Reference Model," *ConneXions,* Volume 4, No. 12, December 1990.

[2] Malamud, C., "DECnet/OSI Phase V: Real OSI or Only Selected Interfaces?," *ConneXions,* Volume 4, No. 10, October 1990.

[3] Wolter, M., "Fiber Distributed Data Interface (FDDI)—A Tutorial," *ConneXions,* Volume 4, No. 10, October 1990.

[4] Kozel, E., "The Cisco/DEC/NTI/StrataCom Frame Relay Specification," *ConneXions,* Volume 5, No. 3, March 1991.

[5] Hughes, L., & Starliper, S., "Switched Multimegabit Data Service (SMDS), *ConneXions,* Volume 4, No. 10, October 1990.

[6] Dern, D., "Commercial IP Providers establish CIX gateway," *ConneXions,* Volume 5, No. 7, July 1991.

[7] Blackshaw, R., "Components of OSI: Integrated Services Digital Networks (ISDN)," *ConneXions,* Volume 3, No. 4, April 1989.

NICK LIPPIS is a Principal Consultant at Strategic Networks Consulting, Inc. He specializes in LAN-to-WAN interconnection architectures and implementations. His MCI Mail address is LIPPIS. (lippis@mcimail.com)

JAMES HERMAN is a Principal at Northeast Consulting Resources, Inc., a consulting firm focused on strategic management and information technologies.

ANSI X3T9.5 Update: Future FDDI Standards

by Mark S. Wolter, National Semiconductor Corporation

FDDI Standards

The ANSI FDDI committee, X3T9.5, has been making progress in terms of both refinements and enhancements to the FDDI set of standards.

SMT

Since the release to Letter Ballot of revision 6.2 of the *Station Management* (SMT) document in May 1990, there have been hundreds of comments that needed to be resolved before forwarding the document. Over ninety percent of these comments have been resolved, most of which reflected clarifications and minor modifications to the specification. Some of the major changes to the specification include the description of managed FDDI objects and the conformance of this description to the ASN.1 standard. These modifications will include a standard way of describing the multiple MAC and Path objects within the MIB for variations configurations found in concentrators and other complex stations. This standardization allows for network management facilities that will monitor and control all conforming FDDI stations. One example of this is a standard method of describing reconfigurable stations such as concentrators with options for inserting additional ports without re-initializing an FDDI network. Also included will be the clarification of standards versus options for management capabilities such as the use of *Parameter Management Frames* (PMFs), and a clarification as to the way PMFs will be used. Another portion of the SMT specification that was previously left open was the method for dynamic allocation of synchronous frame services. A proposal has been made as to the implementation of this feature which includes the use of an additional reserved SMT group address and *Resource Allocation Frames* (RAFs), which were already defined for this purpose but the method of use was never described.

The current SMT specification and implementations conforming to it allow for the operation of a fully interoperable network. Parameters that are undergoing changes are mostly related to the management capabilities of an FDDI network, and do not effect the operation of the network itself.

The coordinated release and integration of future SMT implementations is recognized as an important task. The *SMT Developers Forum,* an independent group of vendors, has taken on the task to produce an agreement for introducing the latest version of SMT before it becomes a standard, and to then gain experience with the latest enhancements.

Twisted Pair PMD

Since the first record attendance ad-hoc meeting a year ago, there have been many proposals presented for a standard copper-based twisted pair specification for an FDDI alternative *Physical Layer, Medium Dependent* (PMD) solution known as TP-PMD. The reason for this interest is not only to reduce the costs of the PMD components, but to greatly reduce the costs involved with the installation of an FDDI network by using an already installed transmission medium. The proposals for a copper-based twisted pair specification can be grouped into three basic categories related to the technique used for compensating for transmission distortion. In time, the ANSI committee will complete a specification derived from one or more of these techniques.

continued on next page

FDDI Standards Update *(continued)*

The first category uses a post compensation technique, which accounts for the inherent distortion by recognizing the changes in the transmitted signal due to DC-bias and frequency related impedances, compensating for it. This method has been proven to work on Type 1 cable (used for Token Ring 802.5), shielded twisted pair up to 100m, but will not work on unshielded twisted pair wires. It is implemented using discrete circuitry that may not be possible to integrate into silicon.

The second category uses a pre-compensation technique similar to the technique used for 10Base-T Ethernet 802.3. Using this technique, a signal is "pre-distorted" to compensate for the effects of the transmission medium. Since the clock encoding used in FDDI, unlike Ethernet, contains a DC-bias, this technique is more complex than 10Base-T, but still feasible. This method will work on shielded and unshielded Data-grade twisted pair (commonly installed for telephones today, but higher grade than older telephone wire installations). It might be possible that this technique will be compatible with the post compensation technique, assuming certain compromises are made.

The third category uses a transmission technique known as *PR4,* a form of partial response coding. This method actually reduces the frequency required for data transmission by using three-level wave forms, and reduces the susceptibility to noise by scrambling the data to "whiten" the spectral components of the transmission frequency. This complicates the transmitter and receiver functions, and require a redefined PMD for transmitted wave forms, but conforms more directly to the transmission characteristics of unshielded twisted pair, including telephone wire.

Alternative Fiber

Another alternative transmission medium currently under consideration is the use of a low cost fiber PMD (LCF-PMD). The issues being addressed by this working group involve the option of using 200 micron fiber versus the current 62.5 micron standard fiber, and the move toward a lower cost connector. This group is currently in an investigation mode.

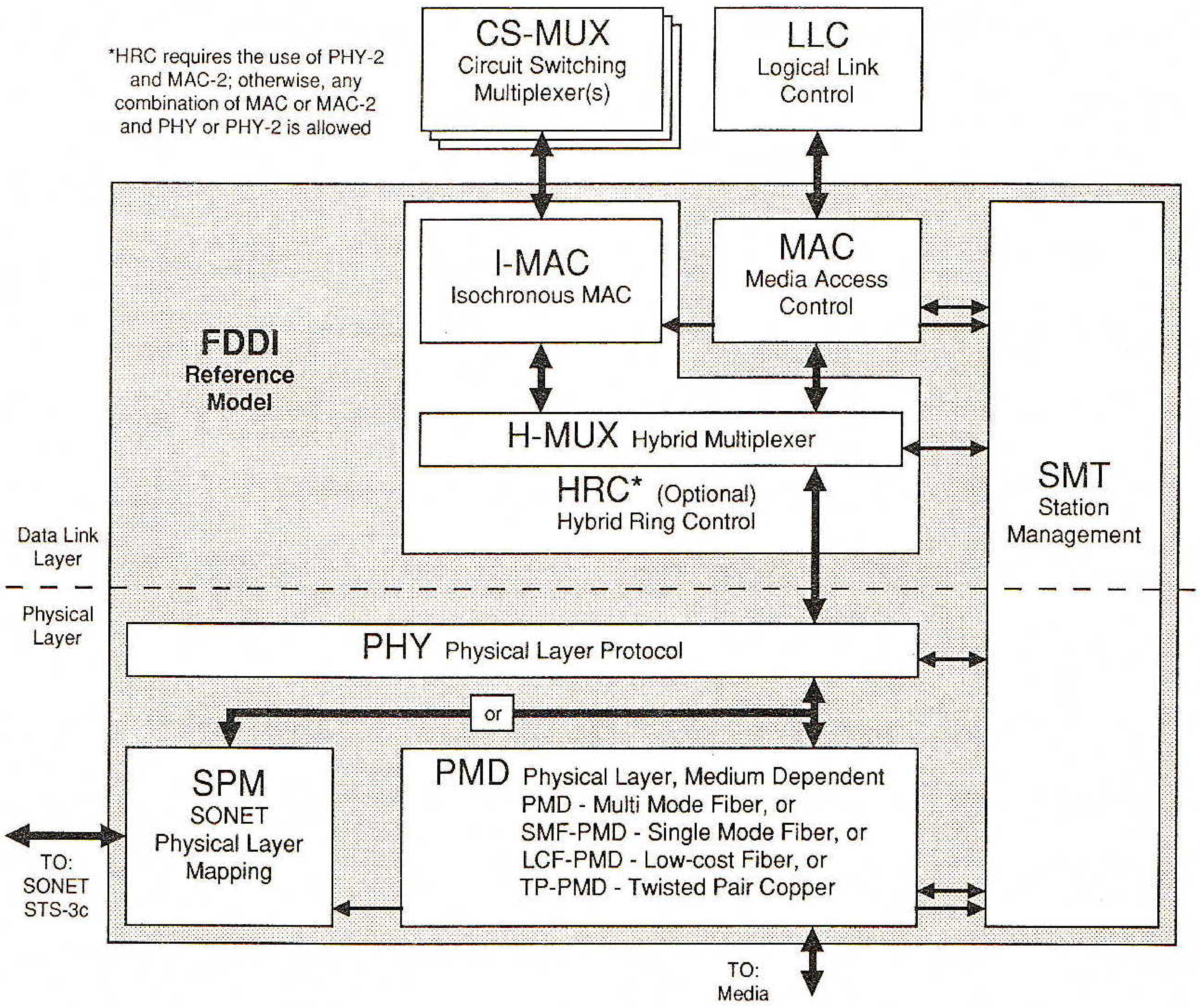
*HRC requires the use of PHY-2 and MAC-2; otherwise, any combination of MAC or MAC-2 and PHY or PHY-2 is allowed

Figure 1: FDDI Standards Structure

Also under specification is the use of single mode fiber PMD (SMF-PMD). This fiber optic cable plant and transceivers are actually more expensive, but allow for longer distances between active connections. This is accomplished because the transmission of a pulse of monochromatic light from a laser diode through single mode fiber has less dispersion per unit of distance than that of an LED diode pulse on multi-mode fiber. This specification uses proven technology and is cost effective for long cable runs versus using active repeating equipment.

SONET mapping

As still another alternative for the Physical Layer, Medium Dependent (PMD) layer of the FDDI standard, a significant amount of interest has surfaced to map the FDDI data transmissions into a SONET payload. This means that in the design of an FDDI station, a SONET network connection would be substituted in place of the standard FDDI PMD. One possible change that may affect higher layers is a change in the 125MHz clock frequency, in order to match the inherent frequencies of a SONET transmission. A first draft of this standard is currently being reviewed.

FDDI–2

The ANSI committee has been working on the specifications for FDDI-2, an alternative network providing integrated services for a fiber optic LAN. The integrated services refers to the support of both an isochronous, circuit-switched protocol as well as a packet-switched protocol commonly used for data communications, such as FDDI. The actual protocol is a hybrid of these two traditional protocols, providing fixed-length cells that can be reserved for circuit switching multiplexers used by the isochronous services, but allowing unassigned cells to be allocated to packet use.

The isochronous services will be used to carry a regular flow of data through a reserved bandwidth channel, or circuit, between two participating stations. To set up this circuit, a procedure similar to making a phone call must be followed. The initiating station's application must know the address of the target station's application, and a circuit must be reserved. During the use of this circuit, addressing and packet length information are not needed to transmit each cell of data.

FDDI Follow On

With the continuing evolution in networking technologies, the FDDI committee will continue its effort to provide additional standards that meet the technologies as they become available. The *FDDI Follow On* (FFO) working group is currently evaluating several networking protocols that will provide additional services required by such applications as multi-media and WAN support. Some possibilities for enhancements include: the incorporation of error correcting codes, provisions for security, multiple data rates in the LAN, seamless connection with the public network, and cycle structures similar to that of FDDI-2 and SONET.

Conformance testing

The objective of conformance testing is to define precisely what requirements must be tested to prove adherence to the FDDI specification. Currently, the *Conformance Testing Committee* has developed a complete set of test suites for the MAC, PHY, and PMD entities for this purpose. This document is currently out for review, and requires a painstaking process to complete due to the detailed review required. Once SMT is finalized, a conformance test suite will be written for SMT as well.

MARK WOLTER is an applications section head for the Advanced Communications Group at National Semiconductor and participates on the ANSI FDDI committee. He has written papers and presented seminars for several systems conferences. He has experience in the design of both FDDI networking systems and ICs. He can be reached as wolter@berlioz.nsc.com.

Developments in SMDS

by Padma Krishnaswamy and Mehmet Ulema, Bellcore

Introduction

Readers of this publication have already been introduced to SMDS, and to the rationales that spurred its development. [1] In this account, we provide a more detailed service description, and an update on the significant progress SMDS has made in the last year.

SMDS development is geared to a three phase SMDS "schedule" with each phase incorporating increased functional capabilities. SMDS, Phase 1, is targeted at individual *Local Access and Transport Areas* (LATAs) where high demand exists. IEEE 802.6-based customer access over DS-1 and DS-3 interfaces, and appropriate levels of operations and network management support are part of this phase.

SMDS Phase 2 is targeted for availability in late 1992-early 1993, and will support greater connectivity and more extensive customer network management features. In this phase, it will be possible for subscribers directly connected to *Local Exchange Carriers* (LECs) to communicate with subscribers in other LATAs via an *Inter-exchange Carrier* (IXC). The proposed LEC service in support of inter-LATA SMDS is called *Exchange Access SMDS* [2], and will be offered to Interexchange Carriers. With the increased Customer Network Management support [3], SMDS customers in this phase will be able to make use of features that will assist them in managing their extended network, in particular the subnetwork consisting of SMDS. In the latter portion of this phase (late 1993-1994), the carriers will have the capability to interconnect switches from different suppliers. The *Inter Switching System Interface* (ISSI) [4] will be used to interconnect these switches within an intra-LATA network.

The third phase of SMDS is intended to coincide with the introduction of BISDN, which will support a variety of existing and new services including SMDS, voice, and video. This phase is planned to start in 1995. As BISDN is introduced, the service layer of the SMDS protocol suite will remain unchanged, thus causing minimal impact to the SMDS subscriber. SMDS Phase 3, will be characterized by the addition of the BISDN interfaces (i.e., User-Network Interface and Network Node Interface) being specified in the national and international standards bodies.

The service

SMDS was developed by Bellcore and the Bell Operating Companies to address customer needs to interconnect computers, workstations and multi-megabit Local Area Networks (LANs). SMDS is designed to be a high performance switched service to provide connectivity, and extend LAN-like performance, across a wide geographical area. In particular, the switched nature of the service will provide the flexibility for any-to-any communication with cost savings over the corresponding leased line topologies. A major driving factor in the service design is to facilitate its incorporation into subscribers' networks with minimal modifications to the internetworking equipment already in use. Accordingly, the network providing SMDS is designed to assume the role of a subnetwork in a customer's communications architecture, as shown in Figure 1. This illustrates that SMDS provides a *Medium Access Control* (MAC) level service.

Basic features

SMDS will transfer variable-length user data, which can be as long as 9188 octets. [5] Each SMDS datagram encapsulates the user data (including protocols above the MAC layer) and contains source and destination addresses that are inserted at the sender *Computer Communications Equipment* (CCE), and delivered to the receiver along with the user data.

The addresses identify customer interfaces to SMDS and are structured according to the E.164 numbering plan. SMDS permits multiple addresses to be assigned to a single customer interface. Group addressed packet transport, similar to the multicasting feature of a typical LAN environment, is a key feature of SMDS. A single group addressed datagram is sent into the network which then delivers it to the set of individual interfaces identified by the group address.

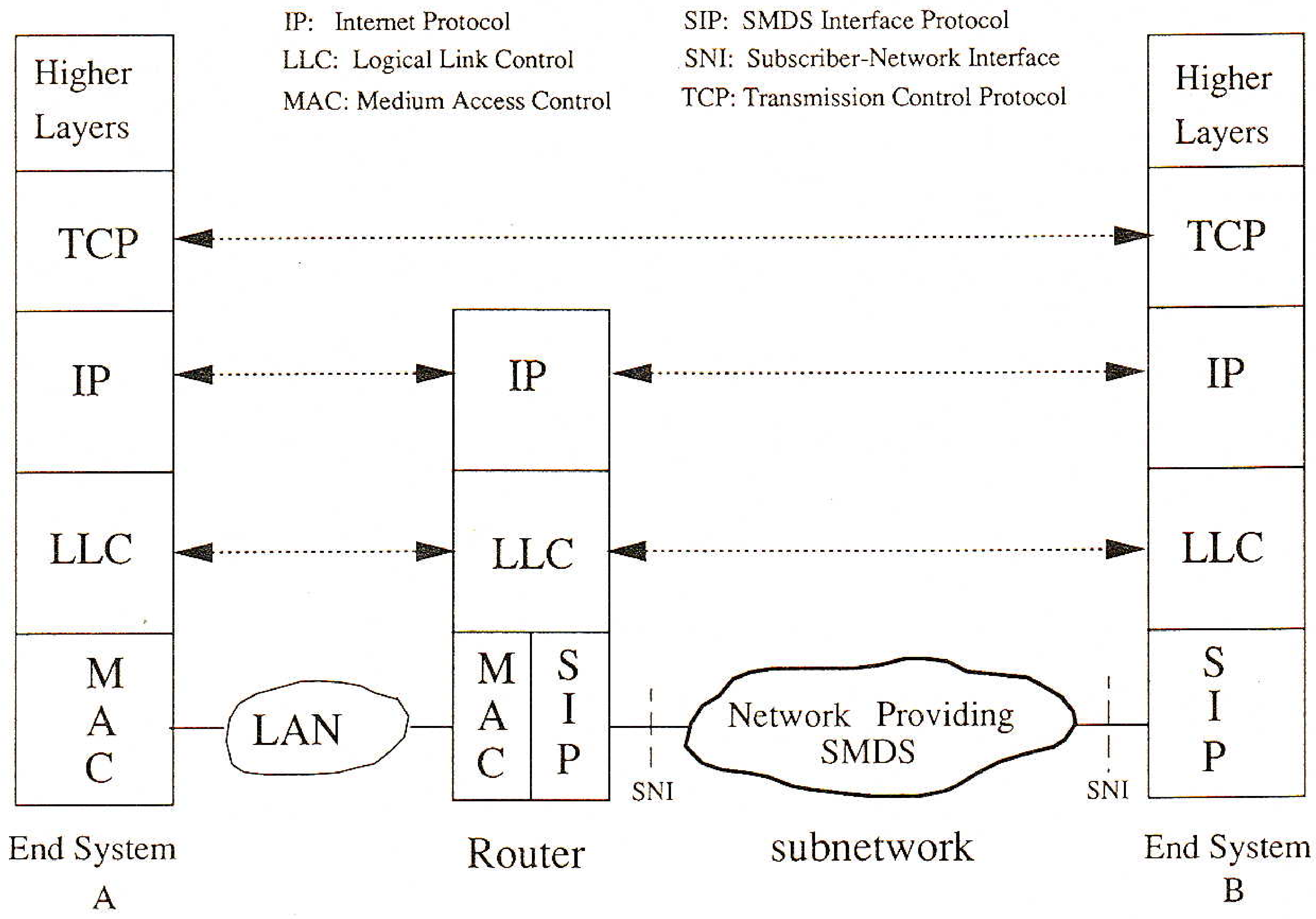

Figure 1: Role of a network supporting SMDS in the customer's communications architecture (using Internet protocols)

SMDS allows customers to set up private virtual networks via address screening features. Both source and destination address screening are offered, with the screening lists being either inclusive or exclusive. The SMDS network validates all source addresses to help ensure that the specified source address is legitimate on the interface from which the packet was sent. An additional feature, available with SMDS access at DS-3 rates and above, accommodates varying application and equipment capabilities at the sending end by superimposing a range of bandwidths on the same 44.7 Mbps physical layer data rate. These bandwidth ranges are called "Access Classes," and effect different traffic characteristics by placing limits on the level of data allowed to flow across the sending access interface. This is distinct from the actual data rate on the physical layer on the access line, which is the standard DS-3 rate of 44.7 Mbps. This is done transparently to the user by SMDS equipment in the network. The Access Classes provided are 1.2 Mbps, 4 Mbps, 10 Mbps, 16 Mbps, 25 Mbps, and 34 Mbps.

Inter-LATA service

A significant factor in maximizing the utility of SMDS is the ability to provide the service over long distance, or, in the U.S.A., across LATA boundaries. This requires the participation of inter-exchange carriers, as well as the specification of the appropriate interface and service. This service specification is referred to as *Exchange Access SMDS* (XA-SMDS), and will be offered in the U.S.A. by the LECs to the IXCs in order to facilitate inter-LATA subscriber communication. For both the basic, or intra-LATA, and the inter-LATA service, the end-user's access to the network would occur across the same *Subscriber-Network Interface* (SNI).

continued on next page

Developments in SMDS *(continued)*

Individually addressed datagrams from the end user are transmitted via XA-SMDS from an SNI to an *Inter Carrier Interface* (ICI), or from an ICI to the destination SNI. The IXC that the subscriber wishes to use may be preselected or explicitly specified in the packet (to override the preselected IXC). As in the case of the basic service, XA-SMDS also supports the source address validation feature for the originating traffic to ensure that the source address is one of those assigned to the SNI. XA-SMDS also supports the transport of group-addressed packets from a single source to multiple inter-LATA destinations. Additionally, XA-SMDS supports end-user blocking which permits an IXC to request an LEC network to block all traffic from an SNI destined for the IXC.

CNM

In addition to the basic service, networks supporting SMDS will offer *Customer Network Management* (CNM) capabilities to let customers access and manipulate network management information pertinent to their SMDS subnet. The CNM service plays a major complementary role to SMDS. In keeping with one of the main tenets of SMDS development, as much synergy as possible has been maintained with existing parallels for the management of data transport services over LANs and LAN network management systems that currently exist and are in operation.

The approach taken to develop the CNM features began with the identification of the information required for effective network management. This falls into the following four broad categories: configuration management, performance management, usage management, and fault management. In each category were incorporated information "items" or capabilities such as:

- *Configuration Management:* access screening and group addressing specifics for the SNI; contact personnel, and equipment location
- *Performance management:* SIP error counts
- *Usage Management:* usage counts recorded by the switch for each SNI
- *Fault Management:* Failure notification, test initiation privileges

The CNM feature suite permits this information to be accessed and interpreted in several ways. Example feature primitives are "Receive Event Notifications," "Request Event Notifications," "Retrieve Subscription Profile," "Retrieve CNM Information," "Retrieve Performance Information," and "Retrieve Usage formation."

The subscriber can gain access to CNM information in three ways: by the use of Management Application Protocol exchange mechanisms (initially SNMP, with the use of CMIP under study), by terminal access, or by requesting it in the form of "hard copy." In the last case, the customer is sent reports containing CNM information generated by the LEC. In the scenario where SNMP is used between the LEC and the customer to exchange management information, the Network Management station is on the customer premises, and the LEC SNMP agent resides on the LEC premises.

Interfaces and protocols

In the following paragraphs, we present brief descriptions of the interfaces and protocols that are defined for an SMDS network. Figure 2 shows these interfaces.

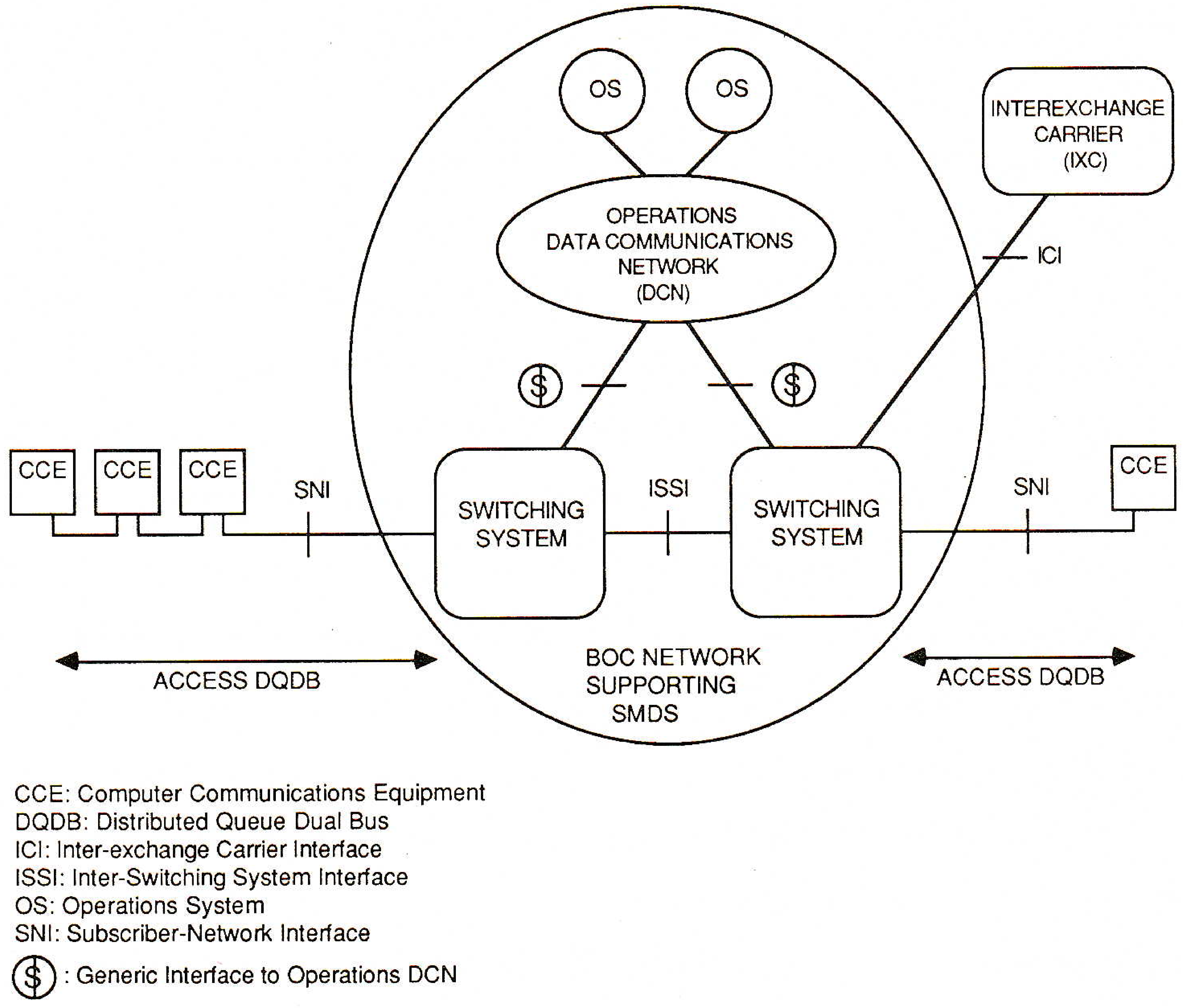
OS
OS
INTEREXCHANGE CARRIER (IXC)
OPERATIONS DATA COMMUNICATIONS NETWORK (DCN)
ICI
$
$
CCE
CCE
CCE
SNI
SWITCHING SYSTEM
ISSI
SWITCHING SYSTEM
SNI
CCE
ACCESS DQDB
BOC NETWORK SUPPORTING SMDS
ACCESS DQDB
CCE: Computer Communications Equipment
DQDB: Distributed Queue Dual Bus
ICI: Inter-exchange Carrier Interface
ISSI: Inter-Switching System Interface
OS: Operations System
SNI: Subscriber-Network Interface
$: Generic Interface to Operations DCN

Figure 2: Interfaces of a network supporting SMDS

The *SMDS Interface Protocol* (SIP) defines how computer communications equipment on customer premises accesses the network supporting SMDS across the SNI. As a service concept, considerable emphasis is placed on SMDS being independent of, and transparent to, the underlying network technology. However, in the interest of open interfaces, the SIP for the initial service offering is based on the connectionless MAC standard [7] developed by the IEEE 802.6 standards body.

SMDS is designed to be offered over the standard telecom data rates of DS-1 (1.544 Mbps) and DS-3 (44.7 Mbps), and, where it is deployed, the newer SONET STS-3c rate of 155 Mbps. The other supported rates are those in the European telecommunications hierarchy of 2.048 and 34 Mbps including the *Synchronous Digital Hierarchy* (SDH) 155 Mbps. The specifications for the U.S. rates are being dealt with by the IEEE 802.6 standards group; for the rest, including the SDH version of the 155 Mbps rate, standardization is being undertaken by the *European Telecommunications Standards Institute* (ETSI). Specifications for DS-1 and DS-3 rates are complete; and the SONET STS-3c PLCP standard is near completion. The ETSI standards are also close to finalization.

SMDS subscribers will interface with the network supporting SMDS at the *Subscriber Network Interface* (SNI). [5, 6] The SNI is the demarcation point between the customer's equipment and the network's equipment. The access path between the subscriber and the network is dedicated to the equipment of a single customer to establish a private and secure network service. The customer may have one or more pieces of computer communications equipment attached to the access path.

continued on next page

Developments in SMDS *(continued)*

The SIP consists of three protocol levels providing addressing, framing, error detection, and physical transport functions. The principal features of the service are realized by the highest level (SIP Level 3) of the access protocol. The SIP Level 3 *Protocol Data Unit* (PDU) is the datagram responsible for providing the connectionless service. It contains the SMDS addressing information and encapsulates the user data. This is contained in the Level 3 PDU's variable length information field (up to 9188 octets). The remainder of the L3_PDU format consists of various length and tag fields to allow fragmentation into, or reassembly from L2_PDUs.

Level 2 of the SIP also provides some functions associated with the segmentation and reassembly of the variable length L3_PDU. The L2_PDU has a fixed-length (53-octet) format containing a 44-octet information payload (Segmentation Unit). The remaining octets contain fields that provide bit error detection mechanisms and framing capabilities.

The SIP Level 1 provides the physical interface to the digital network. It has two sublayers: the *Transmission System* sublayer which defines the characteristics of, and method of attachment to, the transmission link, and the *Physical Layer Convergence Protocol* sublayer, which maps the SIP Level 1 control information and L2_PDUs into a format suitable for transmission on the digital bit stream.

The *Interexchange Carrier Interface* (ICI) provides the connecting link between LEC and IXC networks. The ICI connects a switch in a LEC network to a switch in an IXC network and serves as a point of service termination between the LEC and the IXC. There may be multiple ICIs between an LEC network and an IXC network. The ICI Protocol (ICIP) describes the procedures and functions that are needed on both sides of the ICI to communicate. The ICIP, based on the IEEE 802.6 standard, is a three-level protocol suite designed after the SIP model.

The *Inter-Switching System Interface* (ISSI) will be used to interconnect two switches within a single LATA. It provides a non-proprietary, standard interface to interconnect switches from different vendors, potentially based on different technologies. In addition to its principal function of packet transport, the ISSI also supports OSPF-based routing [8] and congestion management and the necessary LEC operations functions. [4] This protocol (ISSIP) also has a three-level architecture that is similar to the layering structure of the SIP.

Service and technology trials

There is considerable SMDS activity in the data communications industry, in the form of technology trials and product announcements by equipment manufacturers (both switch and CCE vendors) introducing SMDS support on their products.

In the U.S.A., the various LECs are actively pursuing SMDS offerings. In addition, IXCs are showing increasing interest in SMDS and are engaged in discussions with some of the LECs about service and technology trials.

Bell Atlantic's trial with Temple University in Philadelphia, which began in October 1990, concluded successfully in April 1991. In this instance, LANs and a data center on the main campus were linked with LANs at the Health and Science campus. The trial technology operated at DS-3 rates, and was based on a preliminary version of IEEE 802.6 equipment from QPSX. From the results of this trial, and other market research, it was concluded that there was a high level of demand for an SMDS-type service.

Pacific Bell ran extensive trials that included several labs at Stanford University as well as some companies in the Bay Area. Sun Microsystems, Apple Computers, Cisco Systems, Tandem Computers, the US Geological Survey, Pacific Gas and Electric, as well as an internal customer at Hayward, were involved. This set of trials, which began last November, concluded this June. The switching equipment used was from AT&T. SMDS was provided at DS-1 rates. The three switches used in the trial were located in San Ramon, CA, and on the premises of Stanford University. Most of the trial customers used SMDS for file transfer over interconnected LANs. Stanford University used SMDS to transfer images between the school and two locations over ten miles apart. The trial was considered a success by the participants, and led to Pacific Bell planning to offer SMDS in early 1992. For the fall of 1991, Pacific Bell is also planning a joint trial with GTE in the Los Angeles area.

BellSouth began an internal SMDS trial connecting company LANs to link service order negotiating systems at several sites. This was in addition to an existing SMDS concept trial using AT&T equipment. BellSouth also embarked on a collaborative research effort with IBM using the IBM *PARIS* technology, to test SMDS and other broadband services.

Southwestern Bell commenced an internal trial of SMDS last year which is being expanded in 1991. They demonstrated a teleradiology application using SMDS at *Supercomm '91,* in Houston. A single X-ray picture was viewed simultaneously on two different monitors using the high speed switching capability of SMDS.

Also at *Supercomm '91,* US WEST announced their intent to initiate the three-year *Communications Programs for Advanced Switched Services* (COMPASS), which will include a trial of broadband technologies, with SMDS being one of the services. SMDS will be used both for medical imaging applications and LAN interconnection. This will be a joint technology trial, with the other participants being Siemens-Stromberg-Carlson, Fujitsu, and AT&T.

NYNEX expanded its internal SMDS technical trials, and continued to be active in fostering networking applications development. As an example, a trial was recently announced in the Boston area, that features collaboration on distributed application development with New England Telephone, four hospitals, and a local publisher.

Ameritech's internal SMDS tests are under way, with trials involving external customers scheduled to begin in the second half of 1991.

Both long distance and Independent Local Exchange Carrier companies have expressed strong interest in SMDS. MCI publicly announced their support for SMDS at the June 1991 International Communications Association conference. WilTel stated their intent to support SMDS in literature that is publicly available. In addition, GTE, the largest independent LEC, is to begin an internal trial of SMDS, that will last from mid '91 to early '92. GTE also announced that it will be participating in a joint trial with Pacific Bell in the Los Angeles LATA.

SMDS in Europe

Interest in SMDS in Europe is high, as indicated by the large number of PTT trial plan announcements, as well as the participation of many European equipment vendors in standards activities. The chief standardization vehicle for the European market is the ETSI MAN working group.

continued on next page

Developments in SMDS *(continued)*

This group is working on modifications of SMDS for the European telecommunications and networking environment, mostly in the MAN context.

In 1991, trials are planned, or are underway, in England, Denmark, Germany, Holland, Switzerland, Austria, Italy, and Sweden. Trials are expected in 1992 in Finland, Norway, Iceland, Ireland, Belgium and Luxembourg.

An important milestone was achieved in December 1990 with the finalization of the IEEE 802.6-1990 standard. As mentioned earlier, the IEEE connectionless MAC interface is the basis for the SIP.

Standards and industry groups

Equally important was the formation of an industry consortium called the *SMDS Interest Group* (SIG). The group is intended to further SMDS implementation. It includes service providers, equipment vendors, and users committed to the advancement of SMDS worldwide. The SIG works to enhance public understanding of SMDS and its applications, and to support interoperability among different networks and equipment. Activities in the SIG are organized into Working Groups in four categories: the Technical Working Group, the Network Management Group, the Intercarrier Issues Group, and the Publicity and Information Resource Group. One noteworthy example of the kinds of initiative fostered in the SIG is the move to generate a working specification for an HDLC based interface between an SMDS router and DSU/CSU to carry SMDS traffic. Two independent teams developed this interface, the final shape of which is in the process of being ironed out in the SIG Technical Working Group.

Issues concerning *Simple Network Management Protocol* (SNMP) support for CNM, such as *Management Information Base* (MIB) definitions, and methods to exchange IP packets over SMDS are being worked in the *Internet Engineering Task Force* (IETF).

Equipment vendor activities

On the equipment supplier front, leading router manufacturers such as Cisco Systems, Wellfleet Communications, Advanced Computer Communications, Ungerman-Bass, and others, have made SMDS product announcements. In several technology trials, teaming has occurred between specific vendors and the LECs involved.

Some router companies have also formed collaborations with prominent CSU/DSU vendors such as ADC Kentrox, Digital Link, Verilink and NEC to further product and interface development. In addition, Base$_2$ Systems has announced its intent to produce an SMDS integrated circuit device. Switching system suppliers such as AT&T, Alcatel, Fujitsu, Siemens, and NEC, have also announced SMDS support on their products.

INTEROP demonstrations

Cisco Systems, Kentrox, AT&T, BellSouth, NYNEX, Pacific Bell, and Southwestern Bell participated in a successful demonstration at INTEROP 90 that illustrated the power an SMDS solution could bring to the support of high-speed, distributed-processing applications. Remote sites in St. Louis, Atlanta, Cedar Knolls (NJ), White Plains (NY), and several booths on the floor, were connected to a switch in San Ramon (CA) to support nationwide information exchange using SMDS technology.

A demonstration of SMDS is again planned for INTEROP 91: this time, the focus will be on a much broader equipment and application base. It will feature multi-vendor CCE equipment compliant with the IEEE 802.6-1990 standard.

The intent is to illustrate DS-3 rate service, inter-enterprise applications over SMDS, inter-vendor CCE compatibility, and SNMP management for SMDS. Emphasis will be placed on imaging applications such as scientific visualization, CAD/CAM, and multimedia conferencing. The roster of participants registered for the SMDS display is impressive; at this time, there are twenty-seven.

Conclusion

SMDS has come a long way since its inception. Judging by the aggressive clip at which progress has occurred over the last year, it continues to be a major service initiative. It holds forth the promise of an attractive alternative in the search for wider, faster, and better networking. Widespread participation in working through service and equipment related issues has occurred; not merely the telephone companies, but data communications manufacturers, other carriers, and standards bodies have involved themselves in the process.

The service and technology trials have users, carriers, and equipment manufacturers working together towards practical solutions. SMDS fits today's needs, (with deployments in 1992) but also provides a clear evolutionary path to the BISDN of tomorrow. SMDS could usher in a new era in inter-enterprise connectivity. It brings to the field the advantage of the economies of a switched service while offering high speeds, to support high-performance applications.

References

[1] L. Hughes and S. Starliper, "Switched Multimegabit Data Service (SMDS)," *ConneXions,* Volume 4, No. 10, October 1990.

[2] TA-TSV-001060, "Exchange Access SMDS Service Generic Requirements," Bellcore Technical Advisory, Issue 1, Dec. 1990.

[3] TA-TSV-001062, "Generic Requirements for SMDS Customer Network Management Service," Bellcore Technical Advisory, Issue 1, February 1991.

[4] TA-TSV-001059, "Inter-Switching System Interface Generic Requirements in Support of SMDS Service," Bellcore Technical Advisory, Issue 1, December 1990.

[5] TR-TSV-000772, "Generic System Requirements in Support of Switched Multi-megabit Data Service," Bellcore Technical Reference, Issue 1, May 1991.

[6] TA-TSY-000773, "Local Access System Generic Requirements, Objectives, and Interfaces in Support of Switched Multi-megabit Data Service," Bellcore Technical Advisory, Issue 2, March 1990 plus Supplement 1, December 1990.

[7] IEEE Standard 802.6-1990, "Distributed Queue Dual Bus (DQDB) Subnetwork of a Metropolitan Area Network (MAN)."

[8] "The OSPF Specification," RFC 1131.

PADMA KRISHNASWAMY is a Member of the Technical Staff at Bellcore, and works on SMDS. Before joining Bellcore, she worked at the Cornell Information Technology center at Cornell University, Ithaca, New York, where her responsibilities included engineering and maintaining the University data communications networks.

MEHMET ULEMA received his BS (1972) from Istanbul Technical University, and MS (1977) PhD (1980) from Polytechnic University, Brooklyn, NY. For the past two years he has worked for Bellcore mainly on developing congestion management strategies for SMDS. His current interests are protocols, routing and performance analysis of Broadband ISDN and Metropolitan Area Networks. Previously, he was employed by AT&T Bell Labs where he was involved in architecting, designing and analyzing data communications networks based on X.25, Frame Relay, and BISDN technologies. He participated in national/international standards meetings. He was the editor of CCITT's 1988 X.75 Recommendation. He also worked for Hazeltine Corporation, NY, on various software development projects and DARPA's Packet Radio program.

The Service for the Internet Decade

The Multi-Protocol Internet at Delmarva Power
A Case Study

by John K. Scoggin, Delmarva Power & Light

Introduction

Delmarva Power & Light (DP&L) is an investor-owned electric and gas utility serving approximately 360,000 electric customers and 83,000 gas customers on the Delaware-Maryland-Virginia Peninsula. Delmarva places major emphasis on customer service, and has thirteen district and division offices to provide personalized service. The company operates four major power plants on the Peninsula, as well as a number of combustion turbines for peaking service.

The computing environment at DP&L is diverse; an IBM 3090-200J running MVS/ESA and VM/XA provides general business data processing and timesharing services to an SNA network of about 1000 devices. A number of dedicated processors from companies such as Tandem, Hewlett Packard, Digital Equipment Corporation, and Intergraph are utilized for special applications such as Computer-Aided Radio Dispatch, Fuel Dispensing, Meter Reading, Building Management, Automated Mapping, and Direct Load Control. Office automation at Delmarva Power is largely based upon IBM-compatible personal computers running MS-DOS applications. The Banyan *VINES* network operating system is used for providing enhanced services such as mainframe access, common disk storage, electronic mail, and device sharing. Three Control Data Cyber 170 systems are used for energy control applications.

Wide Area Network

In 1987, Delmarva embarked on the construction of a private wide area network based upon digital microwave and fiber optic cable placed on existing electrical transmission towers. Since that time, 200 miles of digital facilities have been placed in service. This private network links most of Delmarva's power plants, division and corporate offices, major district offices, and the Corporate Data Center. This network supports all of the company's voice and data needs, as well as a host of utility-specific requirements such as protective relaying and substation control. (See Figure 1).

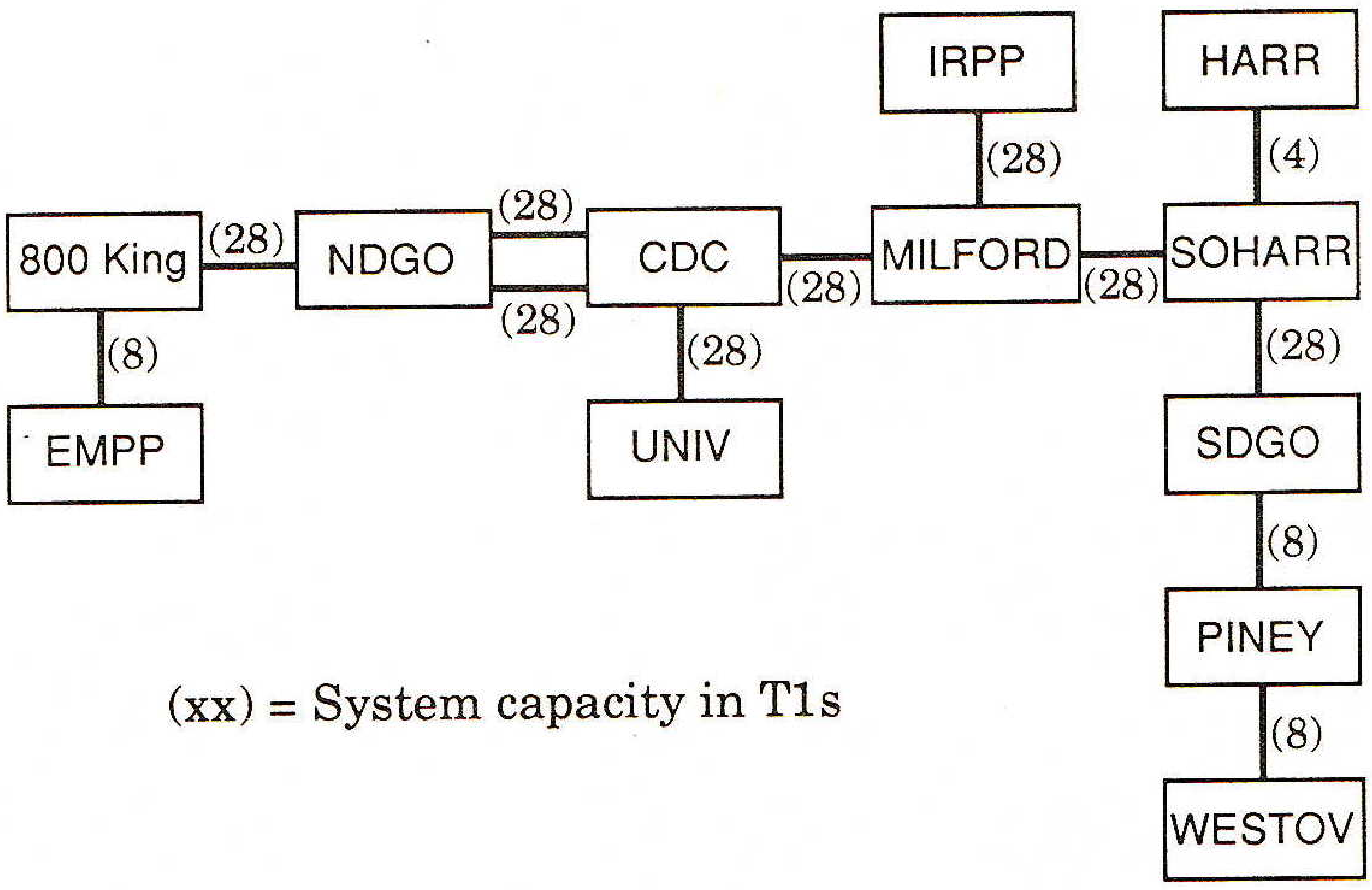

Figure 1: Delmarva Power Digital Transmission Network

The construction of this network significantly altered the economics of data and voice communications at Delmarva. Additional capacity can be gained through modest one-time costs—one T1 from Newark, Delaware to Salisbury, Maryland, a path of 90 miles, costs $1,500 to install on the fiber optic backbone.

Networking Initiative

The rapid influx of mainframe-based on-line applications and a general interest in improving access and information sharing among the diverse computing systems at Delmarva had created a serious problem for the Network Operations section. A tangle of special links between systems and users had become increasingly difficult to support from a hardware and software standpoint. The growth of office automation and distributed systems requests also pointed to the need for a general, vendor-independent network approach.

In the fall of 1988, Delmarva Power participated in a trial of Northern Telecom's *Meridian Data Networking System* (MDNS). MDNS was touted as a general-purpose networking platform providing mainframe connectivity and typical PC LAN services over an X.25 backbone. System performance was poor, however, and significant differences developed between Northern's product development plans and Delmarva's needs. The MDNS trial was terminated in the spring of 1989, and the product was discontinued shortly thereafter.

In the summer of 1989, Delmarva Power's Network Operations section convened a *Network Direction Task Force* to establish a data communications strategy for the early '90s. Representatives from various sections within the Information Systems Group (User Services, System Development, and Network Operations) and selected network users convened with consultants from KPMG Peat Marwick for one week to consider design alternatives.

A week of brainstorming produced a plan to install a TCP/IP-based internet, connecting computers and workstations on IEEE 802.3 local area networks with the existing wideband digital network. It was discovered that nearly all of the systems currently installed had the capability of providing terminal access and file transfer using DOD standard protocols. The largest problem appeared to be the existing SNA 3270 network. Performance requirements appeared to prohibit the incorporation of this network into the new Delmarva internet.

Network management was a major issue; the incorporation of so many vendors' equipment pointed to the need for a management standard. The *Simple Network Management Protocol* (SNMP) was selected because it was the only network management protocol available for commercial use. The use of proprietary network management products would be avoided, unless the required investment in hardware and software were small and the vendor had SNMP capability in development.

Implementation

Vendor selections for the test network were made in the fall of 1989. The IEEE 802.3 networks were based upon Cabletron's Multi-Media Access Centers, largely based on the availability of twisted-pair and fiber optic interfaces, as well as the strong management capabilities of the product. Routers from Wellfleet were selected due to their ability to be managed using SNMP and to perform a bridging function. The Banyan VINES servers were equipped with the TCP/IP protocol option for server-to-server communication. Banyan also provides a server-based copy of Telnet for terminal access and FTP for file transfer in the MS-DOS environment.

It was quickly found that a "pure" TCP/IP solution was not immediately feasible. The Intergraph CAD system used *Xerox Network System* (XNS) for access to a VAX system. Cabletron's 802.3 hubs used an IEEE 802.1-compliant packet for network management.

continued on next page

The Internet at Delmarva Power *(continued)*

Although migration to TCP/IP was in the plan, it was not possible at that point. The Wellfleet routers' bridging function was engaged and the pilot network was operational. The Intergraph CAD network and Banyan VINES networks were the first corporate internet systems.

Integration of the IBM 3090 was more involved. A McDATA 6100E Ethernet Processor was installed with IBM's VM and MVS TCP/IP software. After a short debugging period, file transfer and Telnet terminal access were achieved.

SNA

The integration of the SNA network on the corporate internet was made possible with an 802.3 LAN-attached 3270 cluster controller, made by McDATA, the 4174-44R. The existing McDATA 6100E was enhanced with another LAN attachment and a software module. The MVS VTAM tables were changed to show the addition of a token ring-attached 3174 cluster controller to the channel. The 6100E bridges this SNA traffic onto the internet where they are received by the 4174-44R controller. (See Figure 2). No changes to applications or terminal equipment were necessary!

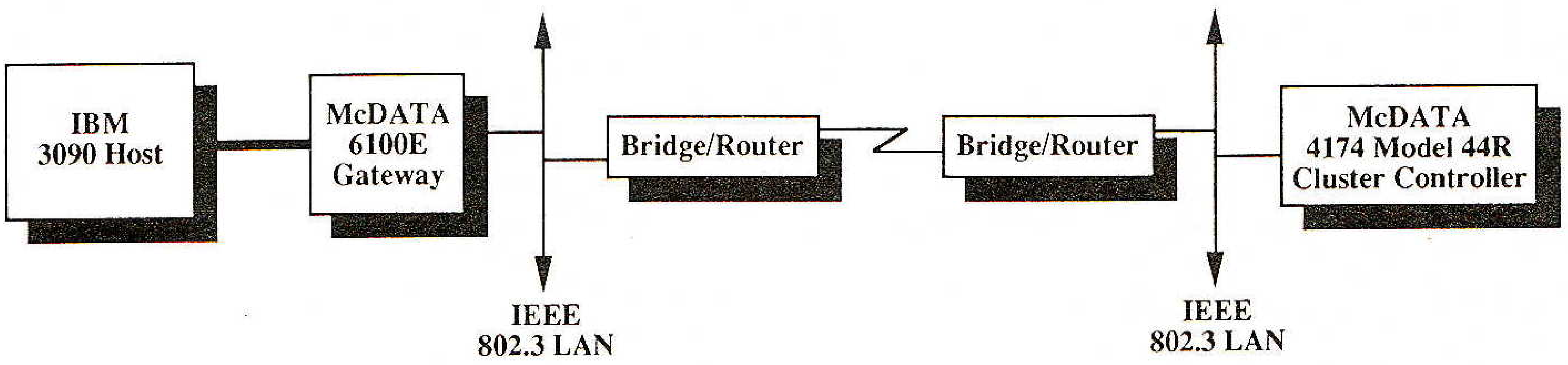

Figure 2: SNA Traffic Integration

User impact

The installation of a corporate internet with high-speed links provided immediate benefit to our CAD and VINES users. These users were previously linked via 56 kilobit/second lines; the new 448 and 672 kilobit/second lines have greatly improved response times. The first user of the LAN-attached 3270 equipment noted sub-second response time on IBM mainframe applications.

Until the introduction of the corporate internet, we had limited penetration of LANs and office automation applications. With the advent of sub-second mainframe access and vastly improved data access times, the Banyan VINES network is growing rapidly. MIPS and IBM RS/6000 systems were easily incorporated into the network in a few hours; conversion from mainframe-based timesharing was eased by the availability of high-speed file transfer and print service. Our first distributed database application, a *Power Plant Operations and Performance System,* is in the design stage. A number of new applications utilizing cooperative processing (MVS-VM-PC) are in development.

The reliability of the internet routing and transport hardware has been remarkable. In the first twelve months of operation there have been no unscheduled network outages of over a few minutes, other than those caused by wide-area network failures. The short outages were caused by software changes made by Network Operations which necessitated router reboots.

Network management

The use of SNMP has been steadily increasing throughout the Internet community. The installation of Cabletron's *SPECTRUM* network management system will permit management of all major network components; LAN hubs, routers, terminal servers, and VINES servers.

Work is proceeding in-house to link our wide-area network equipment (T1 multiplexors and fiber optic terminals) into this framework. By the end of 1992, it is hoped that all of our major network assets will be manageable through the SPECTRUM system. Management of host systems will be more difficult—although IBM has announced SNMP capability for *NetView,* its usefulness in our network is yet to be determined.

The one remaining difficulty is the establishment and monitoring of service levels in this new environment. Service levels on the IBM-based SNA network were measured with our Dynatech *PRISM* Performance Measurement System. Availability and response time objectives have been set for the SNA 3270 network, but have not yet been defined for the internet environment. Delmarva is developing a monitoring system for the Banyan VINES environment with the goal of measuring user-perceived service levels on LAN-based applications. This system will also be part of the SNMP-based management architecture. (See Figure 3).

Finally, the physical simplicity of IEEE 802.3 equipment is simplifying the job of attaching user equipment to the network. Training requirements should be considerably reduced as the old network equipment is replaced with Internet-based products.

Network Management System	Elements
Racal-Milgo CMS400	Modems Low-speed (<56 Kbps) muxes
Racal-Milgo CMS6000	T1 Muxes Master for CMS400
DPL SNMP Proxy Agent	Rockwell 3x50 Lightwave System
DPL SNMP Proxy Agent	Pulsecom Datalok 10 Facilities alarms
Cabletron SPECTRUM	Ethernet hubs Routers VINES servers Terminal servers
Delmarva DIALS (voice notification for above systems)	All
IBM NetView	SNA Network Elements

Figure 3: Network Management Architecture

Conclusion

The conversion to a multi-protocol internet is expected to be complete by 1995. Obsolete equipment is being replaced with the newer technology as required by new applications or obsolescence. The availability of a ubiquitous high-speed network will accelerate the development of distributed applications and permit the rapid sharing of information and ideas with all employees.

JOHN K. SCOGGIN, Jr. has been active in the data processing and telecommunications arena for 17 years. His career has spanned a number of positions including both engineering and business applications programming, systems programming, and telecommunications. His current job responsibilities include the design, operation, and maintenance of a regional voice/data network based upon approximately 200 route-miles of privately-owned fiber optic and digital microwave systems and the operation of the User Help Desk. John is currently an Adjunct Assistant Professor in the Computer and Information Sciences Department of Goldey-Beacom College, teaching courses in Operating Systems and Data Communications. Off the job, John is a volunteer firefighter/EMT, a member of the Delaware State Fire Police, and the Deputy Officer of the Delaware Division of Emergency Planning and Operations. He can be reached as: `scoggin@delmarva.com`.

INTEROP 91 Fall will present this case study in the *"Emergence of the Common Corporate Internet,"* session S28, at 3:30 pm on Thursday October 10, and the Corporate Internet BOF the same evening at 9:15 pm.

Applications and Techniques for LAN Monitoring

by Jeffrey Mogul, Digital Equipment Corporation,
Western Research Laboratory

What is LAN monitoring?

Most network communication starts, ends, or stays on a *Local Area Network* (LAN). LANs are wonderful, except when something goes wrong, which is often. Finding out what has gone wrong can be extremely difficult, because the source of the problem might be spread out among multiple components of the LAN. [The original paper on Ethernet was subtitled *Distributed Packet Switching for Local Computer Networks;* the Ethernet has been called a "distributed single point of failure."]

Just as electrical engineers have tools such as oscilloscopes and logic analyzers to help monitor electrical systems, network engineers have tools for monitoring LANs. This article is an overview of the various uses of LAN monitoring, and a look at some of the existing and potential tools.

I define *LAN monitoring* as

- the collecting of packets on a broadcast LAN
- by a host other than their destination
- without the active involvement of source and destination hosts
- and presentation of packet information in a useful form.

LAN monitoring is different from other kinds of network management techniques because it is passive; that is, it does not require any support from the LAN or its hosts, and it does not affect the traffic being monitored. The distinction is important, because it means that you can use LAN monitoring without having to install special software on any of the monitored hosts, and without perturbing their operation (which might change the nature of the situation being analyzed). It also means that in most cases LAN monitoring is the only way to discover aggregate behavior of the network, such as performance statistics and precise timing relationships.

LAN monitoring and network management protocols (such as SNMP) are thus complementary approaches to the problem of network management. Management protocols concentrate on individual hosts, require agent implementations on those hosts, and don't work unless both the hosts and the network are at least minimally functional. LAN monitoring concentrates on the LAN as a whole, and can be useful when the host in question (or even the entire LAN) is completely broken. Of course, management protocols are necessary when you need to control a network host or see its internal state, instead of simply watching its network transmissions.

Applications

My definition of LAN monitoring leaves room for a broad variety of applications. In this article, I will describe the most common kinds of applications, but there are many others, some yet to be invented. One nice consequence of the passive nature of LAN monitoring is that you can build a new monitoring application without having to get cooperation from anyone else. You don't need to wait for the IETF to approve a new MIB, or for the vendors to implement it, or for users to install the new software release.

Traffic analysis

If you are responsible for managing a LAN, you often need to know how much traffic is being carried and where it is coming from. Traffic analysis aids in capacity planning (do I need a faster LAN?), in topology planning (do I need to split my LAN using a router or bridge?), and in finding busy hosts (maybe a host is using the LAN inefficiently).

Traffic analysis is concerned with aggregate statistics of packet flows. Network management protocols can provide rudimentary traffic analysis, by allowing a management application to aggregate the packet counts maintained at the hosts on the LAN, but this at best provides a long-term average value for the network load (and may seriously underestimate it if a host is malfunctioning). Since network traffic is usually quite bursty, the long-term load average is not particularly interesting. When people are using interactive protocols such as Telnet, overloads that last for more than a fraction of a second may be annoying.

A LAN monitor, by watching every packet that goes by, can calculate the load average over arbitrarily short intervals. Typically, an interval of one second is short enough to portray the peaks that annoy users, but long enough to ignore the unimportant burstiness of the traffic.

Load averages can be displayed by graphing load versus time (Figure 1); the resulting curve shows not only what the medium-term average is, but how bursty the short-term average is. A load monitor application might also maintain tables of statistical values, such as the peak and average loads for a variety of averaging periods.

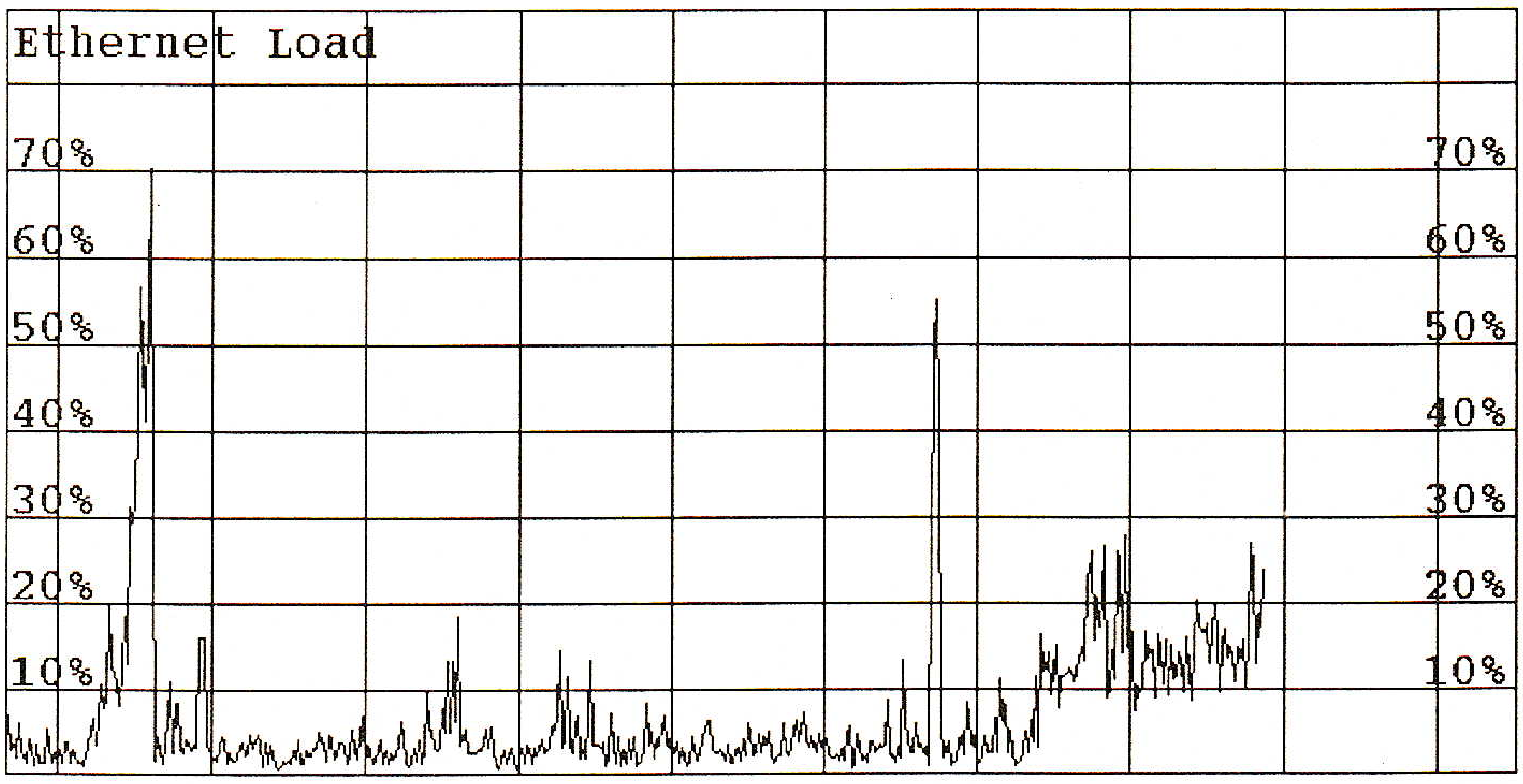

Figure 1: Graph of Ethernet load vs. time

If bursts in the short-term average load stays too high (say, above 60% of the channel capacity) for significant periods, it might be time to reorganize the network. One approach is to find a way of splitting the hosts on the LAN into two new LANs, divided by a router or bridge, so that most of the communication need not cross the division. A LAN monitor can produce a matrix showing how many packets are being exchanged by each possible pair of hosts on the LAN, from which one can calculate (at least in principle) an assignment of hosts to subnets that minimizes cross-router traffic. (This may not always be practical; the traffic patterns may not be stable, or there might be other considerations that govern the assignment of hosts to subnets.)

continued on next page

Techniques for LAN Monitoring *(continued)*

If the load is too high, it may be that one or more hosts are simply operating incorrectly. For example, a host may be "jabbering" (sending out lots of useless packets), or something might be repeatedly trying to retry a failed operation. A LAN monitor can identify which of the hosts on a network are sending the most traffic (Figure 2), and you can then try to figure out if that host is misbehaving.

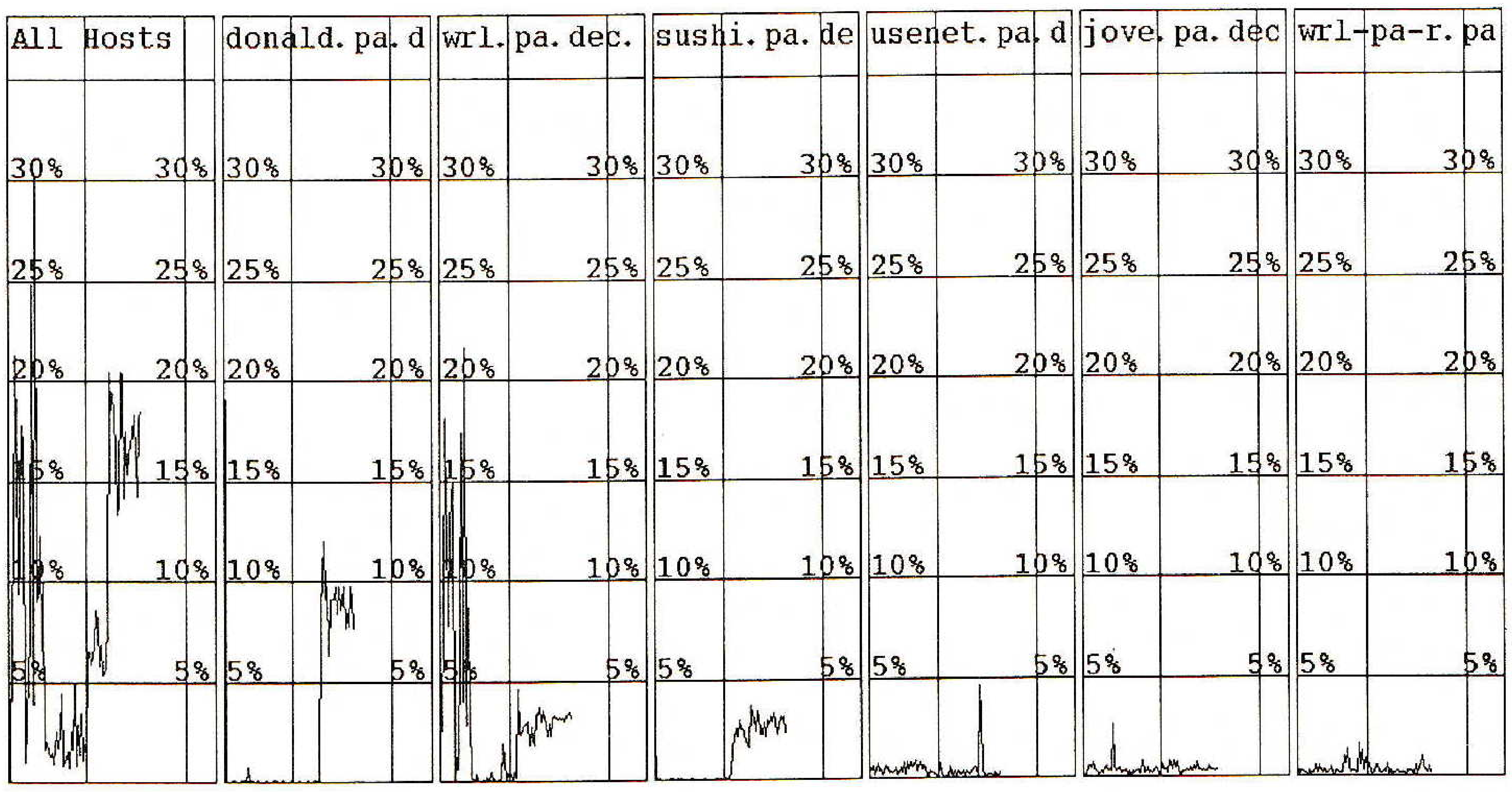

Figure 2: Graphs of Ethernet load broken down by source host

Traffic analysis can also be restricted to a specific protocol type or service. This may be helpful in discovering the sources of high load. For example, it might be useful to know that 99% of the traffic on the network comes from NFS.

In some organizations, it may be necessary to account for a host's network usage on the basis of actual traffic generated, rather than simply dividing the cost up equally. A LAN monitor can be used to count the number and size of packets sent by each host, without requiring those hosts to waste their own effort maintaining this information.

Tracing

When you have to track down a bug, statistical information is sometimes insufficient. You need to be able to look at the contents of individual packets to see what is going wrong. To do this, you need a "packet trace," that is, a sequenced display of some subset of the packets on your LAN.

Tracing tools help find incorrectly configured hosts, incorrect protocol implementations, and sources of bad packets. Since a complex communications system might be broken in any of a number of places, a trace is invaluable in discovering just what part of the system is malfunctioning. For example, if a host is unable to boot over the network, a trace of packets to and from that host can show if it is (for example) successfully sending out BOOTP packets, but then sending TFTP requests to the wrong server.

A tracing tool has two important functions. First, it must allow you to specify exactly the right subset of packets. It could take several minutes or hours to capture the ones you are looking for, and during that time millions of uninteresting packets might go by. It is impossible to manually sift for needles in such a large haystack, so the tracing tool must do it for you. It should let you specify filtering predicates in a natural, high-level form.

The tool must also present the packets in a useful form. Normally, one is only interested in the packet headers, not the entire packet contents, but since protocols may be layered arbitrarily deeply this means that a good tracing tool is able to interpret many different protocols. The results are usually displayed in a cryptic textual format. Verbose formats might seem easier to read, but actually make it too hard to pick out the important information (especially in a long trace).

For example, the trace below (made by the *tcpdump* program) shows an entire SMTP (TCP/IP electronic mail) conversation between two hosts named "dog" and "cat":

```
14:43:28.406 dog.2583 > cat.smtp: S 6904:6904(0) win 4096 <mss 1024>
14:43:28.413 cat.smtp > dog.2583: S 5076:5076(0) ack 6905 win 4096 <mss 1024>
14:43:28.417 dog.2583 > cat.smtp: . ack 1 win 4096
14:43:29.421 cat.smtp > dog.2583: P 1:82(81) ack 1 win 4096
14:43:29.429 dog.2583 > cat.smtp: P 1:22(21) ack 82 win 4096
14:43:29.437 cat.smtp > dog.2583: P 82:148(66) ack 22 win 4096
14:43:29.445 dog.2583 > cat.smtp: P 22:40(18) ack 148 win 4096
14:43:29.562 cat.smtp > dog.2583: . ack 40 win 4096
14:43:29.913 cat.smtp > dog.2583: P 148:174(26) ack 40 win 4096
14:43:29.921 dog.2583 > cat.smtp: P 40:79(39) ack 174 win 4096
14:43:29.960 cat.smtp > dog.2583: . ack 79 win 4096
14:43:30.101 cat.smtp > dog.2583: P 174:226(52) ack 79 win 4096
14:43:30.109 dog.2583 > cat.smtp: P 79:84(5) ack 226 win 4096
14:43:30.160 cat.smtp > dog.2583: . ack 84 win 4096
14:43:30.206 cat.smtp > dog.2583: P 226:276(50) ack 84 win 4096
14:43:30.222 dog.2583 > cat.smtp: . ack 276 win 4096
14:43:30.226 dog.2583 > cat.smtp: P 84:294(210) ack 276 win 4096
14:43:30.363 cat.smtp > dog.2583: . ack 294 win 4096
14:43:30.367 dog.2583 > cat.smtp: P 294:296(2) ack 276 win 4096
14:43:30.562 cat.smtp > dog.2583: . ack 296 win 4096
14:43:30.613 cat.smtp > dog.2583: P 276:284(8) ack 296 win 4096
14:43:30.621 dog.2583 > cat.smtp: P 296:301(5) ack 284 win 4096
14:43:30.761 cat.smtp > dog.2583: . ack 301 win 4091
14:43:31.132 cat.smtp > dog.2583: P 284:326(42) ack 301 win 4096
14:43:31.136 dog.2583 > cat.smtp: F 301:301(0) ack 326 win 4096
14:43:31.136 cat.smtp > dog.2583: F 326:326(0) ack 301 win 4096
14:43:31.140 cat.smtp > dog.2583: F 326:326(0) ack 302 win 4096
14:43:31.140 dog.2583 > cat.smtp: F 301:301(0) ack 327 win 4096
14:43:31.144 cat.smtp > dog.2583: . ack 302 win 4096
14:43:31.144 dog.2583 > cat.smtp: . ack 327 win 4096
```

The first column is a timestamp, and the next two columns (separated by ">") show the packet source and destination (host-name.port). The rest of the line expresses the interesting parts of the TCP header, including (not in order) the Flags field ("S" means SYN, "P" means PUSH, and "F" means FIN), the TCP options, the window size, the number of data bytes (in parentheses), and the various sequence number fields (expressed as an offset relative to the initial sequence numbers). Someone who understands the TCP protocol can learn to read this trace quite easily.

Packet traces that include precise timing information (in this trace, the timestamps are accurate to about 4 msec) allow one to debug performance problems as well as correctness problems. Since well-designed network protocols tend to hide the effects of low-level failures (such as lost packets or poor implementation design), the only clue one might get about low-level malfunctions is inadequate performance. In this trace, one can see a delay of about 1 second between the third and fourth packets; this reflects the time it took for the SMTP server to create a new process and initialize its data structures.

continued on next page

Techniques for LAN Monitoring *(continued)*

Research

In addition to applications in network management and debugging, LAN monitoring is a useful way to gain insights into the performance and traffic patterns of real networks. Such data as this are essential in conducting research on computer networking, because while simulations and theoretical studies can be quite interesting, to be useful they must be based on realistic models of present or future networks. Studies based on monitoring production-environment LANs tell us what kinds of traffic patterns actually arise in practice.

Since LAN monitoring is passive, and avoids perturbing the system being observed, in many cases it is the only way to get precise information about the dynamic behavior of real networks. In principle, one could perhaps obtain the same kind of information using a network management protocol such as SNMP...but in practice, the extra packet traffic and host CPU loading would alter the timing relationships and invalidate the data.

Several examples of monitor-based research are worth mentioning. One study by Raj Jain and Shawn Routhier [3] showed that packets on an Ethernet tended to be bursty and highly correlated with respect to source and destination hosts. Simulation studies often assume that packet transmissions are random and unpredictable; Jain and Routhier's result implies that such studies could be misguided.

Van Jacobson's breakthrough work on congestion avoidance and control [2] was concerned with the behavior of TCP connections over wide-area networks, but he obtained important relevant information by monitoring the connections as they transited a convenient LAN. Van converted the output of *tcpdump* to graphs of sequence number versus time; these graphs allowed him to observe the unfortunate response of early TCP implementations in the presence of congestion. Figure 3 shows a plot of sequence number versus time for part of a TCP transfer over a moderately congested path. Unfortunately, I could not find a system that has not been upgraded to include Van's changes, so the figure shows an almost boringly smooth curve. You can see a few places where retransmissions occurred, interrupting the steady increase in sequence number.

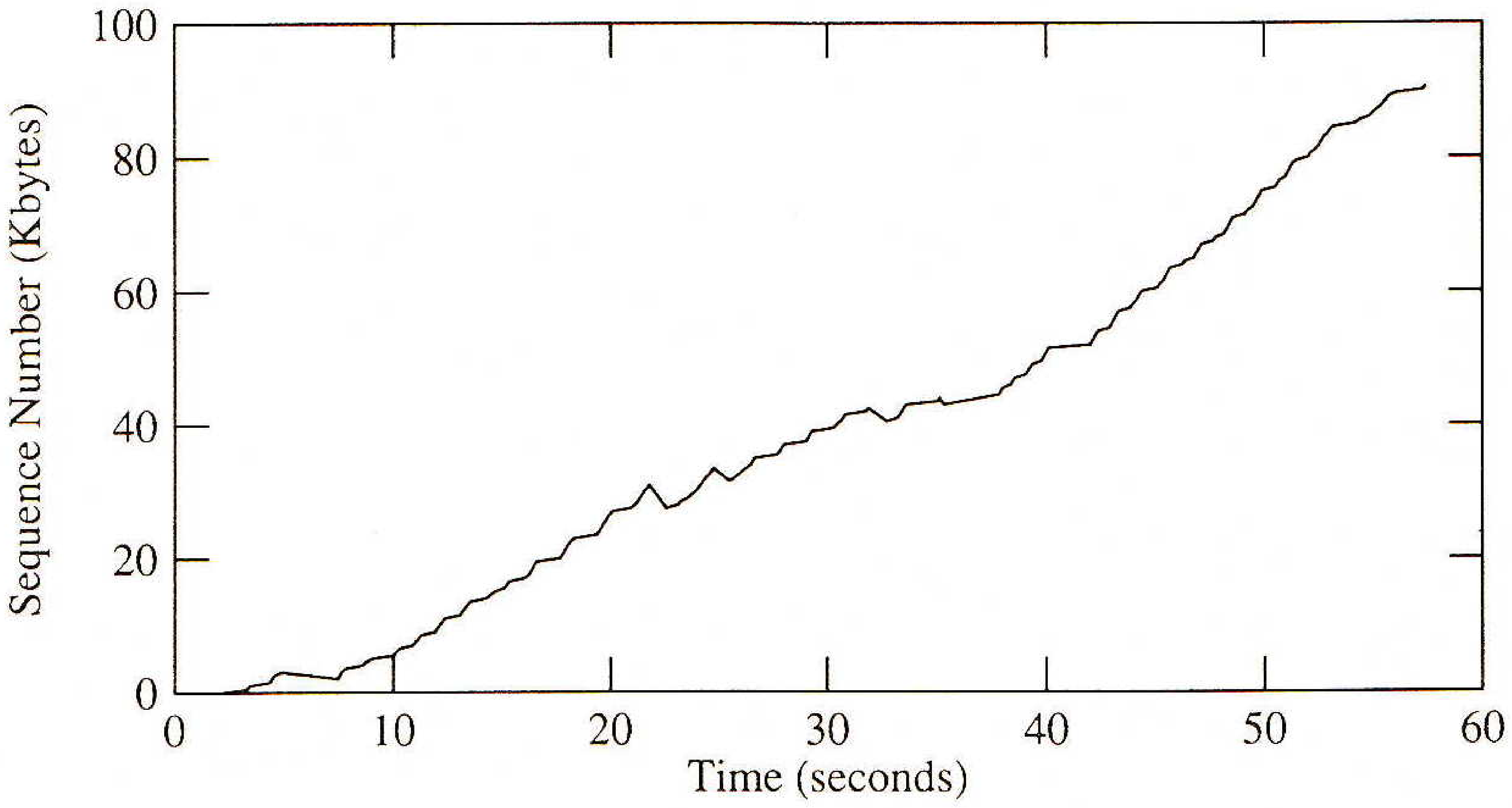

Figure 3: TCP sequence number versus time for an FTP transfer

Abuse

A passive LAN monitor is a great spy tool. Without anyone knowing what you are up to, you can (with not much programming effort) discover what people are typing on their terminals, storing on their file servers, sending in their mail messages, and typing for their passwords. With a little more work, you can obtain access to their files without their permission.

Encryption is the only complete solution to this problem. By using cryptographically-based authentication protocols, one avoids sending passwords (or similarly dangerous information) in cleartext over the network. By encrypting data in mail, remote terminal, and file service protocols, one prevents people from invading one's privacy.

Since encryption is not yet ubiquitous (far from it), passive monitoring is still very much a danger. There is not much one can do to prevent it (almost any PC or workstation can be set up as a LAN monitor), so it is important to take other precautions (such as segregating security-critical information on LANs physically separate from potential spies). It is best to create a social and administrative atmosphere that discourages inappropriate monitoring, since determined spies nearly always bypass technical protections.

Tools and Techniques

Because there is such a variety of LAN monitoring applications, the tools and techniques available also vary considerably. This makes it hard to choose the right tools, since no one tool could support all the potential applications. I won't discuss specific commercial products, both because I am not familiar with everything on the market and because I don't have room to cover the ones I do know about without appearing to favor particular products. Of course, some are better than others, and in buying one you should consider what features you most need. You will probably find it necessary to obtain several complementary tools.

Choice of platform

Monitoring a busy high-speed LAN, because it requires processing every packet, places unique demands on systems engineered for much lower packet rates. You have a choice of hardware and software that span a variety of tradeoffs: performance vs. cost, simplicity of use vs. flexibility, etc.

The least expensive approach to network monitoring is to use a PC computer with a normal LAN interface. You might have many of these on your LAN, or you can buy one quite cheaply. If you don't want to continuously monitor your LAN, you might find it convenient to "borrow" a PC whenever you are debugging a transient problem. You also might find it more convenient to carry around a diskette with monitoring software, rather than an independent piece of hardware. Ubiquity and cost are the two main advantages of using PCs; speed (or its absence) is the main disadvantage. Although some PC CPUs are quite powerful, many PC LAN interfaces are not.

You can buy LAN monitoring software for PCs, and there is also some public-domain software available. Often, the software allows you to add your own code, if you need to decode packet formats not supported by the vendor.

At the other end of the cost spectrum, but often with the best possible performance, are dedicated hardware/software systems. A dedicated system is usually engineered to capture packets at the LAN's maximum traffic rate, and although they do have to be carried around they are usually quite portable. They are also easy to use; you don't need to do anything special to get them started, and some provide a large set of function-specific buttons (as on an oscilloscope), which makes the user interface more efficient.

My favorite approach, partly because it is the most convenient for doing research and developing novel applications, is to use a workstation. Workstations typically have high-performance CPUs and LAN interfaces, and they also have great graphics capabilities and good software-development environments.

continued on next page

Techniques for LAN Monitoring *(continued)*

A workstation might also be able to support several LAN-monitoring applications simultaneously, in different windows, which is useful if you need to observe a problem from several points of view. While workstations are not cheap, in many places they are ubiquitous; if you don't need to do full-time monitoring you can borrow one for debugging a problem. Of course, if you want the software to be useful during LAN-wide failures, it should be stored in advance on the workstation's local disk.

Several vendors sell LAN-monitoring software to run on workstations. There are also a number of public-domain programs available via anonymous FTP. Three that I frequently use are:

tcpdump For obtaining and formatting packet traces

NNstat A package for obtaining statistical information on network traffic; the user can specify a wide range of statistical breakdowns

nfswatch For obtaining statistical information on NFS activity; helpful in debugging NFS performance problems

These and other tools are listed in the so-called "NOCTools Catalog," most recently issued as RFC 1147.

One approach becoming popular, especially in large installations, is the use of small, low-cost collection boxes on each subnet of a large LAN complex, combined with a central analysis and display system (usually a workstation). The collection boxes don't provide a user interface, but instead produce summary information which is then transmitted (over the network or over serial links) to the central station. While this may not suffice for all applications, it works well for statistical information and provides an integrated view of an entire network.

Monitor performance issues

One critical issue in choosing a LAN monitor is whether it is capable of capturing enough packets. "Enough" is a vague term, because for some applications (such as measuring the average load) you can drop lots of packets and still get a usefully accurate value. For other applications, such as debugging a performance problem using a tracing tool, you may not be able to afford to miss any of the interesting packets.

There is no single way to measure LAN monitor performance. For statistical monitors, the most interesting value might be the average packet capture rate. If a monitor or tracing tool, however, drops packets only when they arrive in bursts (with no space between packets), it might be more important to know how long a burst can be before packets are dropped. (Some PC LAN interfaces are unable to receive "back-to-back" packets, and are inadequate for many tasks.)

Tracing tools normally buffer the received packets, since a human could not possibly read the headers as fast as they arrive. (Actually, the buffer usually stores only the packet headers, which can save a lot of space.) The size of this capture buffer may determine whether you can gather enough of a trace to solve your problem, especially when the activity being observed lasts for an extended period. Some products are limited to what can be stored in main memory, but at (moderate packet rates) fast workstations are capable of storing arbitrarily long traces on disk files without losing many packets.

If the buffer is big enough, it can be used to postpone some of the other processing (such as aggregation and visualization), which reduces the likelihood that packets will be dropped. Other techniques for avoiding packet loss are to do the filtering as early as possible (on a workstation, in the kernel rather than in user code) and to use clever programming tricks for the performance-critical parts of the code.

Analysis and presentation

No matter what kind of LAN monitor you choose, the system will perform several phases of analysis and presentation. Analysis starts with a filtering phase, optional for statistical monitoring but unavoidable for tracing. The filtering mechanism lets you specify which packets will be analyzed by the following phases.

In a statistical monitor, but not in a tracing tool, the next phase is to aggregate the incoming data. Load-average calculation simply requires keeping track of the number of packets and the sum of their lengths, but more sophisticated statistics require the analyzer to distinguish between different packets before counting them. For example, if you want to know how much of the load on your network is divided between Telnet and *rlogin* connections, the analyzer must parse the packet headers to isolate TCP packets and then categorize them by source or destination port.

A few LAN monitoring tools now provide some automated analysis of either statistics or packet traces. For example, a monitor might be instructed to raise an alarm if the load on the network rises above a threshold chosen to preserve adequate response, or if it drops to zero for an extended period (indicating, perhaps, a shorted LAN cable). Some people have begun to apply artificial intelligence techniques to the analysis of packet traces, to detect anomalous conditions or to aid in performance tuning. [1]

The final phase is visualization: the presentation of information in a form that, if done well, takes maximum advantage of our human ability to interpret visual information. For statistical monitors, this may be as simple as showing tables of numbers, or graphs of load versus time (e.g., Figure 1), or as complex as automatically laid-out maps showing the logical topology of communication. [4]

For tracing tools, the state of the art in presentation is somewhat cruder, but I expect advances to be made especially as automated analysis becomes capable of extracting relationships between packets. Sometimes, you will want to view the same trace in several different ways, so a tracing tool should be able to store large traces offline for future analysis and presentation.

All of these phases depend on layer-specific knowledge. For example, a program cannot filter or aggregate on TCP port numbers without being able to parse TCP packet headers. Automated analysis requires an understanding of the semantics of various header fields, not just their position and size. When a tool doesn't understand a layer, you may be forced to work with numeric values, which are hard to interpret.

Because LAN monitoring can be applied to such a wide range of problems, the user needs to be able to control all of the phases in order to configure a monitor for the task at hand. You need to be able to specify what packets to filter, what statistical aggregates to create, what analyses are interesting, and how you want to view the results.

continued on next page

Techniques for LAN Monitoring *(continued)*

The quality of a LAN monitoring system in large part depends on how complete its protocol suite is (does it understand all the protocols that you need to analyze?) and how flexible it is (can it be configured to solve your novel problems, or is it limited to the applications anticipated by the vendor?).

Integration with management tools

As I wrote at the beginning of this article, LAN monitoring and network management protocols are complementary. Network managers need both in their toolkits. Even better would be a system that integrates the information from LAN monitoring and SNMP, to provide a single station from which to manage an entire network.

Automatic integration of MIB data and LAN-monitor data, to create a new higher-level stream of information, is still in the future. Some progress is being made: there is an IETF working group developing a MIB for LAN monitoring, which would allow remote LAN monitors to be controlled via SNMP. I suspect that by its very nature, the process of choosing a MIB (that is, a standard set of variables) will limit the utility of this approach, since the hallmark of LAN monitoring is the diversity of potential applications. In the long run, I believe that the integration will have to be done much later in the process, because network management protocols were not designed to carry the immense, complex, and unstructured data streams produced by LAN monitors.

Potential improvements

The pace of innovation in LAN monitoring products is such that I often think of a new feature, only to find that someone is already selling it. I expect that we will continue to see improvement and innovation in user interfaces, both in the mechanisms for specifying filters and aggregates, and in visualization of the results.

The area of automated analysis is ripe for future improvements, but it is likely to be hard and slow work. Expert systems technology, which has generally failed to meet the high expectations of a decade ago, does best when applied to domains where the experts have a lot of experience. Since LAN technology is new and constantly evolving, even the experts often don't know how to proceed, and so creating a competent expert system could be difficult.

As the underlying hardware technology continues to scale (CPU speeds and RAM sizes are following a geometric trend) LAN monitor hardware will continue to improve. I expect to see both lower cost (especially as it becomes unnecessary to build special-purpose hardware) and higher performance.

Although inappropriate standardization could stifle innovation, I do believe that there are areas where standards could promote the development of LAN monitoring. For those areas that we understand fairly well, such as the specification of filtering and a set of basic statistical aggregates, some standardization of features and interfaces would help (in the same way that a standard location of gas and brake pedals makes it easy to get into a rented car and drive it away, whereas it always takes me a while to figure out how to reprogram the radio presets in a new car).

Developers of LAN monitoring applications would benefit from standardization of the operating system interfaces that support LAN monitoring. In the past, these interfaces have not been a high priority among operating systems vendors (many vendors still provide no such support), but I think this is changing.

Application developers would also like more control over system behavior. For example, some popular Ethernet interface chips suppress all garbage packets, as a favor to the higher-level software. LAN monitors usually want to see such packets, and the system (hardware and kernel) should make that an option.

Unchartered territory

Although CPU speeds are increasing, so are LAN speeds (albeit not nearly as smoothly). Current systems (workstations and PCs) are able to keep up with 10 Mbit/sec Ethernets, but may not be able to monitor 100 Mbit/sec FDDI networks without dropping packets. Gigabit LANs will clearly require faster systems, and because I/O performance does not scale as easily as CPU speed, it may be hard to produce cost-effective systems able to monitor gigabit LANs without dropping packets.

In the near future, however, these same limitations may prevent our computers from being able to fully utilize gigabit LANs, and so our LAN monitors may not be faced with an excessive traffic rate.

One trend in LAN technology that could cause problems for LAN monitoring is that the most attractive designs for gigabit LANs are mesh-connected, rather than broadcast busses or rings. On a mesh LAN, not only is there no single place to monitor all the traffic, but there is no guarantee that successive packets from one host to another will follow the same path.

One could still do statistical monitoring by careful placement of several "taps." In order to do tracing, however, it might either be necessary to combine the data from multiple taps using highly-synchronized clocks, or to temporarily restrict the LAN topology so that all packets pass through a single, monitored switch. The former approach might be infeasible, and the latter removes the feature that LAN monitoring not perturb the normal operation of the network. In any event, monitoring a mesh-connected LAN will not be as easy as monitoring a broadcast LAN, because it will not be possible to simply use any LAN host as a monitor.

References

[1] Bruce L. Hitson, "Knowledge-Based Monitoring and Control: An Approach to Understanding the Behavior of TCP/IP Network Protocols," In Proc. *SIGCOMM '88 Symposium on Communications Architectures and Protocols,* August 1988.

[2] Van Jacobson, "Congestion Avoidance and Control," In Proc. *SIGCOMM '88 Symposium on Communications Architectures and Protocols,* August 1988.

[3] Raj Jain & Shawn Routhier, "Packet Trains: Measurements and a New Model for Computer Network Traffic," *IEEE Journal on Selected Areas in Communication,* SAC-4(6), September, 1986.

[4] Jeffrey C. Mogul, "Efficient Use Of Workstations for Passive Monitoring of Local Area Networks," In Proc. *SIGCOMM '90 Symposium on Communications Architectures and Protocols,* September 1990.

JEFFREY C. MOGUL received an S.B. from the Massachusetts Institute of Technology in 1979, an M.S. from Stanford University in 1980, and his PhD from the Stanford University Computer Science Department in 1986. Dr. Mogul has been an active participant in the Internet community, and is the author or co-author of several Internet Standards. Since 1986, he has been a researcher at the Digital Equipment Corporation Western Research Laboratory, working on network and operating systems issues for high-performance computer systems. He is a member of ACM, Sigma Xi, the IEEE Computer Society, and CPSR. Address for correspondence: Digital Equipment Corporation Western Research Laboratory, 250 University Avenue, Palo Alto, California, 94301 (mogul@decwrl.dec.com).

Hear more about this topic in S13: *"Tools for Network Monitoring and Control"* at 3:30pm Wednesday, October 9.

The Great IGP Debate—Part One:
IS–IS and Integrated Routing

by Radia Perlman and Ross Callon,
Digital Equipment Corporation

Background

A panel at INTEROP 91 Fall will explore the relative merits of the *Intermediate System to Intermediate System Intra-Domain Routeing Exchange Protocol* (IS–IS) [3, 9] versus the *Open Shortest Path First* routing protocol (OSPF) [8] as routing protocols, and the merits of Integrated Routing versus "Ships in the Night" (S.I.N.). Although these two issues are really orthogonal, in this article we will discuss our position on both issues. We will discuss the advantages IS–IS has as a routing protocol when compared with OSPF, and the advantages of supporting multiple Network Layer protocols with an integrated approach rather than a Ships in the Night approach.

The first portion of the article deals only with IS–IS versus OSPF. It will not attempt to discuss all the differences between the two schemes. Instead, it will discuss only what the authors feel are the differences most likely to be significant. The second portion of the article deals with Integrated Routing versus Ships in the Night.

Similarities

Both protocols are link state protocols. Either one would serve the IP community as a substantial improvement over RIP [7], the current de facto standard for routing IP. [6] Both offer hierarchical routing, variable length subnet masks, multiple types of service, path splitting, and authentication. Either one could replace RIP without affecting the other portions of the IP Network protocol (IP addressing, IP data packet formats, ARP, ICMP, etc.).

The remainder of this section will discuss differences between the protocols. The assumption will be made that the reader is reasonably familiar with both protocols.

OSPF can be used for routing IP traffic. In contrast, IS–IS can be used for routing both IP and OSI *Connectionless Network Layer Protocol* (CLNP) [2]. In order to make a fair comparison, we will compare OSPF and IS–IS for routing of IP traffic only.

Feeding Level 2 information into an area

In IS–IS, level 1 routers only know what is reachable within their own *area*. If a destination address is not located within the area, a level 1 router routes the packet to the nearest level 2 router. In contrast, OSPF feeds information about destinations outside an area into an area. Level 1 routers then choose the level 2 router that gives the best path to the external destination.

Our next generation routing protocols should be designed so that they can scale for use in a global internetwork. There are likely to be at least tens of thousands of addresses reachable outside an area. The IS–IS scheme saves a significant amount of memory in level 1 routers, since they do not need to know information about destinations outside of the area, as well as a significant amount of bandwidth, since this information need not be propagated on links inside the area.

An additional advantage of the IS–IS scheme is that it is possible to predict the memory required by a level 1 router. If an area is planned and deployed and attached to the global internet, the memory required is dependent only upon the size of the area. There is no possibility that at some point in the future the internet will grow so large as to overflow the capacity of the level 1 routers.

The IS–IS scheme increases the separation between the operation of level 1 and level 2 routing. This improves the robustness of the protocol by limiting the potential for propagation of error conditions.

The OSPF scheme gives the possibility of better routes to destinations outside an area. However, we feel better routes will not be a large factor for several reasons: In many environments, the majority of conversations will be between nodes within an area. Also, any bandwidth gained by OSPF, because of utilization of better routes, will be offset by any bandwidth required for knowledge of these routes (i.e., the bandwidth used by informing level 1 routers of level 2 information).

For small and medium sized environments—up to several hundred routers in a single routing domain—it is feasible with either protocol to make most or all routers be level 2 routers (note that hierarchical routing is still useful in this case as it minimizes the spread of level 1 information). In this case the two protocols will calculate equally good routes. For routing domains in which the topology is carefully designed, the level 2 backbone will be well connected and again the quality of the routes calculated will differ very little. Thus OSPF will calculate significantly better routes only for very large routing domains (containing thousands of routers) in which the topology is not planned. However, it is unlikely that a routing domain with thousands of routers and links arranged with no topology planning can be made to work efficiently with either routing protocol.

It is possible to configure an area in OSPF to be a "stub area." In a stub area, information about destinations outside the autonomous system is not fed into the area. In this case OSPF becomes like IS–IS for destinations outside the *Autonomous System* (AS). No bandwidth or memory is required to inform level 1 routers about these destinations, but the routes are the same as used by IS–IS. For destinations outside the area but within the AS, OSPF still propagates this information into an area, even if it has been configured as a stub area.

Encoding efficiency

In IS–IS, the intent is that each router issue a single LSP, containing information about all its neighbors. If the router is both a level 1 and a level 2 router, it will issue two LSPs. One LSP contains the level 1 neighbors and is propagated only within the area. The other LSP contains the level 2 neighbors and is propagated only within the level 2 net. If an LSP is so large that it cannot fit into a single packet, a router breaks its LSP into several fragments, which get independently propagated throughout the network. Each fragment may contain hundreds of destinations.

In OSPF, link state information is reported in *Link State Advertisements* (LSAs). Certain types of link state information are likely to be fairly large. For instance, an AS border router might report thousands or tens of thousands of addresses reachable outside the AS. OSPF requires each destination to be reported in a separate Link State Advertisement. Although many LSAs can be combined into a single packet for transmission purposes, each LSA requires its own header. Overhead information such as sequence number and age must be repeated for each destination. This overhead information takes up memory (for storing it) and bandwidth (for propagating it). Although IS–IS requires the same overhead information, in IS–IS it occurs once for hundreds of destinations, and thus the overhead is amortized over hundreds of destinations.

continued on next page

IS–IS and Integrated Routing *(continued)*

For example, for each IP destination reachable outside an Autonomous System, OSPF requires 36 octets of information. If multiple types of service are supported, OSPF requires an additional 12 octets for each type of service supported, for each destination. Thus if 3 types of service are supported, OSPF requires 60 octets per destination.

In IS–IS information about all 4 potential types of service are always reported, so the overhead is the same regardless of how many types of service are supported. In IS–IS the information required per external destination is 12 octets.

Considering that there are likely to be tens of thousands of destinations reachable outside an AS, the factor of at least 3 (and potentially as much as 6) for storage of level 2 information can be important.

Note that the comparison is for propagating and storing level 2 information. In OSPF, in a non-stub area, level 1 routers have to store all of that information. In IS–IS, level 1 routers do not store or propagate this information.

One potential advantage of OSPF's encoding is that incremental updates are smaller. In OSPF, when a single destination changes, only that single LSA needs to be propagated. In IS–IS, if a single destination changes, the entire affected LSP fragment must be propagated. Whether this is important or not depends on the frequency of changes versus periodic exchange of information. In both schemes, each LSP (or in OSPF terminology, each LSA) must be regenerated by the source router on the order of once per hour. If a very small portion of the link state information changes within an hour, then the savings gained by smaller incremental updates will not be significant.

Managing parameters

The IS–IS protocol is designed so that it is always possible to migrate from one parameter setting to another in a working network, without disrupting network operation. In OSPF, neighbors will refuse to talk unless their parameters are identical. The relevant parameters are:

- *HelloTime* and *DeadTime:* This information informs a neighbor how often a router is going to issue Hello messages, and how long a period without receipt of Hello messages a router should wait before declaring the neighbor down. In IS–IS it is possible to have the HelloTime configured differently at all the different routers on a LAN. Indeed there might even be reasons (besides migrating the LAN from one value to another) for wanting the timers to be different. There might be some routers that are only concentrators for a few end systems. Since there is no other path to those end systems, and those routers are not used for routing to any other destinations, it might be unimportant to notice quickly when those routers are down. To reduce overhead, the HelloTime parameters on those routers might be set to a very large value.

Since in IS–IS it is possible to have different values for HelloTime at each router on the LAN, it is easy to modify these parameters in a working network. In OSPF all routers on the LAN must have the same value for HelloTime and DeadTime. If the network manager decides that the current value is not sufficiently responsive, or is using too much overhead, network operation will be disrupted until every router on the LAN has been reconfigured with the new value.

• *Stub area flag:* IS–IS has no such parameter. In OSPF, all routers in an area must agree on the setting of this flag. If an area is configured with the flag in one setting, and the network manager decides to change the area, then again the area will be disrupted while routers are reconfigured one by one. During reconfiguration routers will refuse to talk to neighbors with the flag at a different setting. If the network manager is operating from one location, the order in which the routers are reconfigured is important, because it is possible for the network manager to become "painted into a corner" where pockets of routers can no longer be reached.

• *Authentication password:* Both OSPF and IS–IS have an optional feature of providing authentication. In OSPF there is a single password. In IS–IS there is a single transmit password, but a set of receive passwords.

In IS–IS, if a LAN is using password P1, and the manager decides to change the password to P2, the manager adds P2 to the set of receive passwords. The manager does this at each router, one at a time. Once this has been done (at all routers on the LAN), the manager changes the transmit password of each router, one at a time, to P2. Once this has been done at each router, P1 can be removed from the set of receive passwords of each router, again one at a time. During this entire process, the network continues to run properly without interruption.

In OSPF neighbors will not communicate unless they have the same password. There is no way to change the password on a LAN without disrupting operation on that LAN while the routers are being reconfigured.

LSP database overload

There are two reasons why the LSP database might become larger than a router was configured for. It might be a temporary situation, which can be caused, for instance, by a change of *Designated Router* on a LAN. Temporarily there might be multiple copies of information and thus temporarily the LSP database might be larger than routers were configured to support. The other reason is that the network has grown larger than a router has been configured to support.

OSPF does not explicitly state what a router should do if it finds it has run out of resources and cannot store link state information. It can crash and automatically reboot; it could crash and wait for manual intervention (which could not be carried out across the network); or it can "fake it," i.e., continue routing on a subset of the routing information. Faking can cause global disruption, since link state routing only works if all routers are making decisions on identical databases. Crashing and rebooting will cause a temporary situation to resolve without human intervention. However, it can cause global disruption in a permanent overload situation, since the link state database will keep changing for all routers as the links to that router cycle. Crashing without rebooting will not cause global disruption, but it can be extremely inconvenient to reconfigure the router when it is not reachable via network management.

IS–IS has mechanisms so that a temporary overload situation will resolve itself without manual intervention, and a permanent overload will not cause global disruption. Also, the IS–IS mechanisms ensure that even a permanent overload will be repairable via network management from any location in the network.

continued on next page

IS–IS and Integrated Routing *(continued)*

The IS–IS mechanism involves a bit in the LSP indicating that a router cannot contain the LSP database. A router that cannot store an LSP sets that bit in its own LSP. When some time has elapsed during which the router is not forced to drop any LSPs, the router clears that bit. A router with the "overloaded" bit set in its LSP is treated, by the other routers, like an end system. Network management packets can be routed to it and accepted from it (as well as any other data packets), but routes are not computed through that router.

An LSP database overload is more likely to occur in OSPF because level 1 routers have to store information about destinations outside the area, and it is very likely that the internet would grow beyond the estimated size at the time some level 1 routers were deployed. When overload does occur, it will be much more difficult to repair with OSPF and be potentially much more disruptive until it is repaired (depending on whether the implementors choose an option that can cause global routing disruption).

Integrated routing vs. Ships in the Night

Integrated routing refers to using a single routing protocol to route multiple Network Layer protocols (for example, Integrated IS–IS may be used to calculate routes for both IP and for CLNP). *Ships in the Night* (S.I.N.) refers to using a different routing protocol for every network layer protocol supported (for example, OSPF might be used for supporting IP, and OSI IS–IS might be used for CLNP).

The issue of whether to use Integrated Routing or Ships in the Night is really completely orthogonal to the issue of OSPF versus IS–IS. However these two issues often become confused.

IS–IS has been modified to support IP as well as CLNP, but OSPF has not been so modified. Thus if the advantages of Integrated Routing are considered important, IS–IS is the only open multi-vendor Integrated Routing protocol available. It is theoretically possible to do Integrated Routing with a routing protocol other than IS–IS. At least today, no protocols other than IS–IS have been explicitly designed and standardized to support multiple Network Layer protocols.

Some people assume that if Ships in the Night is considered the correct approach to supporting both IP and CLNP, that the only choice is IS–IS for CLNP and OSPF for IP. This is not the case. Any routing protocol that supports CLNP can be used in parallel with any routing protocol that supports IP. For example, if a routing domain's administrators decided that they prefer IS–IS to OSPF, but also prefer Ships in the Night to Integrated Routing, then they could use two instances of IS–IS, one for routing of IP only, and a separate instance for routing of OSI only (IS–IS provides a "protocols supported" field, which could be used for demultiplexing the two occurrences of IS–IS, as well as authentication fields which could be used as extra insurance against accidental confusion of the two instances by a mis-configured router).

Management

In this section we give our reasons for supporting Integrated Routing rather than Ships in the Night. Routers require a lot of management, especially for IP. Links have to be assigned costs, addresses, and subnet masks. In the Ships in the Night method of supporting multiple Network Layer protocols, these parameters need to be separately configured for each Network Layer protocol. For *n* protocols, it is *n* times as much work for the network manager. Similarly, when error conditions occur in the network, *n* different routing protocols will respond simultaneously to the same error, resulting in competition for resources and complexity for the network manager.

Integrated routing reduces the management problem to the management of a single routing protocol. Error conditions are responded to by the single protocol. Configuration can be done only once, although those configuration items which are independent for each protocol suite (such as addresses) are configured independently.

Topological considerations

Integrated routing requires the use of a single network topology for support of multiple Network Layer protocols. This includes use of a single common area structure and a single backbone. S.I.N. advocates claim that in some sites it is necessary to configure links separately, because some resources are owned and controlled by the users of one Network Layer protocol, NL1, and other resources are owned and controlled by users of a different protocol, NL2. If a link is owned by the NL1 users, then it can be assigned a small cost for NL1 and a large cost for NL2, and traffic for NL2 is only routed over that link as a last resort. Alternately, the cost for NL2 can be set to infinity, and NL2 traffic will never be routed over that link. This can result in a different topology for each Network Layer protocol supported (although the set of topologies for all these protocols would be overlapped). Such a topology would preclude use of Integrated Routing.

The problem with this approach is the expense. Particularly serious is the cost of additional personnel for managing each routing protocol. Also, since each topology would be different, topology planning becomes much more complex. The result of link failures will be different for each of the multiple Network Layer protocol supported. Similarly, use of different topologies for each protocol greatly complicates the task of capacity planning (planning must be done separately for each topology, although the interactions between the different protocols on shared links must also be considered). In contrast, the use of a single common topology for all Network Layer protocols allows capacity planning to be based on the total overall required capacity.

In some cases links can be owned by different sets of users that happen to use the same network layer protocol. If it is essential to be able to configure link costs so as to control which links are shared, and which links are used solely by a subset of the users, Ships in the Night will not solve the problem. If this is indeed a problem, it can be solved by accounting and cross charging for resources, or by type of service routing. Links can be configured to have different costs for different types of service, and the various user communities can be assigned specific service types they are to use in their data packets.

Bandwidth and memory

When multiple routing protocols are deployed to support multiple Network Layer protocols, the overhead is roughly multiplied by the number of protocols supported. Each protocol will have its own neighbor handshaking procedure. Each protocol will list and propagate router neighbor information independently. With Integrated Routing, the overhead per additional Network protocol supported is very small.

Real-time behavior

Routing protocols are real time systems. A router that, for example, transmits routing packets too slowly or fails to receive packets from its neighbors quickly enough may cause its neighboring routers to decide that it is down and bring down their links to the misbehaving router (which may, of course, cause the load on the router to be reduced, which allows the router to operate correctly, which causes the other routers to decide that it has recovered, thereby allowing the router to become overloaded again, etc). In order to guarantee correct operation of a router during congested operation, it is necessary to ensure that certain real time constraints are met. This is easier to do when only one routing protocol is being used.

continued on next page

IS–IS and Integrated Routing *(continued)*

Development resources

Routing protocols are complicated and subject to failure. It is difficult to design distributed algorithms that do not have hidden failure scenarios in arcane circumstances. A good example of a routing protocol with a design flaw is the collapse of the ARPANET. [10]

Even if the routing protocol is designed correctly, there is the possibility that an implementor might implement it incorrectly, either because of a software bug or because of a misunderstanding of the specification. We suspect most of the readers of this article have come across an incident in which an implementation of a correct protocol had a bug. Even if the routing protocol is designed correctly and implemented correctly, there is the possibility that it can be configured improperly.

It is irresponsible to require networking vendors to understand, implement, and debug more routing protocols than are necessary. If a routing protocol exists, then it should be used, unless there are overwhelming reasons why the world would be better understanding, implementing, and debugging a new one.

In order to support CLNP, the world will need IS–IS. IS–IS can also support IP. If IS–IS is used for supporting IP, then the world will only need to understand, implement and debug IS–IS. If OSPF is used for supporting IP, then the world will need to understand, implement, and debug both OSPF and IS–IS.

References

[1] Ross Callon, "Integrated Routing for Multi-Protocol TCP/IP–OSI Environments,", Proceedings of the *Second Joint European Networking Conference,* forthcoming.

[2] "Protocol for Providing the Connectionless-Mode Network Service," ISO 8473, March 1987.

[3] "Intermediate System to Intermediate System Intra-Domain Routeing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-Mode Network Service (ISO 8473)," ISO DIS 10589, November 1990.

[4] Radia Perlman, "A Comparison Between Two Routing Protocols: OSPF and IS–IS," *IEEE Networks,* forthcoming.

[5] Jon Postel, "Internet Control Message Protocol," RFC 792.

[6] Jon Postel, "Internet Protocol," RFC 791.

[7] Chuck Hedrick, "Routing Information Protocol," RFC 1058.

[8] John Moy, "OSPF specification," RFC 1131.

[9] Ross Callon, "Use of OSI IS–IS for Routing in TCP/IP and Dual Environments," RFC 1195.

[10] Eric C. Rosen, "Vulnerabilities of Network Control Protocols: An Example," *Computer Communications Review,* July 1981.

RADIA PERLMAN has been designing bridge and Network Layer protocols at Digital Equipment Corp. for the last 10 years. She designed the *Spanning Tree Algorithm* which is an essential component of bridges, and designed most of the protocols involved in ES–IS and IS–IS. She holds a Ph.D. in Computer Science from MIT.

ROSS W. CALLON is with the Distributed Systems Architecture group at Digital Equipment Corporation in Littleton, Massachusetts. He is working on OSI–TCP/IP Interoperability issues, and is the architect for the Integrated IS–IS protocol. Mr. Callon is also the co-area director of the OSI Integration area of the IETF, and is the chair of the IETF IS–IS working group. Mr. Callon received his B.Sc in Mathematics from the Massachusetts Institute of Technology, and his M.Sc in Operations Research from Stanford University.

The Great IGP Debate—Part Two:
The Open Shortest Path First (OSPF) Routing Protocol

by Milo S. Medin, NASA Science Internet Office

Abstract

The *Open Shortest Path First* (OSPF) routing protocol is quickly becoming the protocol of choice for routing inside of large IP internets. This article will describe the basic workings of the protocol and compare and contrast it to other IP routing protocols. OSPF's unique features will be described and examples given of its performance and use in real operational networks.

Introduction

The OSPF protocol is a routing protocol for the DARPA Internet Protocol (IP) network layer. It was defined by a working group of the Internet Engineering Task Force, which is the development and engineering organization for the Transmission Control Protocol/ Internet Protocol (TCP/IP) suite. It was developed because of the lack of a really robust and full featured routing protocol for use inside *Autonomous Systems* (ASs). Because it is designed to be used inside ASs, and not between ASs, it is called an *Interior Gateway Protocol* (IGP). The OSPF protocol is designed to be a full featured and robust protocol, capable of being used inside IP ASs, of small to very large size, and has features to reduce routing overhead, quicken network convergence time, increase security and ease of management, and be able to scale effectively in large environments. OSPF was designed specifically to support the IP protocol, and be optimized for use in the operational Internet, to solve real world problems in real world environments.

Algorithms

Network routing protocols are based on *routing algorithms* that are really independent of the actual subnetwork protocol being routed. The two major ones used by protocol designers are the distance-vector algorithm and the link-state algorithm. An elementary understanding of each is critical to a discussion of the OSPF protocol and how it compares to other network routing protocols. The goal of both algorithms is to route a packet from one point in the network to another point in the network through some set of intermediate routers without "looping," the situation where a data packet may be forwarded across the same link a number of times.

Distance-vector protocols

The *distance-vector* algorithm or "old ARPANET" style routing algorithm, is one of the oldest routing algorithms in use. The basic concept behind distance-vector protocols is that each router sends its routing table to each of its neighbors, and that they in turn merge this routing table with their own routing tables, and then transmit the merged table to each of their neighbors. A router is considered a neighbor of another router if they have a direct network link between them. The routers usually have some sort of metric in their table, which may just be a hop count, in which case the neighbor originating the route in the first place would normally assign a metric of 1 to the route, and each neighbor would then add a metric of 1 to the incoming route's metric and add the route to its table. When the router hears about the route from multiple places, it only replaces its route if it hears the route from another router with a smaller metric than the metric that's currently in its own routing table.

Distance-vector protocols may use a number of different metrics. Many have other features to speed propagation of routes and avoid looping or shorten the time spent in loops. The basic distance vector algorithm is fairly simple to understand, and also fairly easy to code.

continued on next page

OSPF *(continued)*

It is the algorithm used in the original ARPANET, the mother of all modern packet switching networks. It is also the protocol that the RIP, Hello, and IGRP protocols are based on. Very few new routing protocols are distance-vector based.

Link-state protocols

In contrast to distance-vector protocols, *link-state* protocols are based on each router in the system learning the topology of the system, and then building a routing table based on the known topology. In link-state protocols, the routers do not send each other routing tables, but rather information about the links that the router has to other routers. In fact, since all routers in the system need to hear this information, the link information is "flooded" through the entire system of routers. The routers do this by sending the link information to each of their neighbors, and then those neighbors sending it to each of their neighbors, until everyone in the system hears it. Once a router in the system has the information about the links in the network, they can build a tree, with themselves at the root, and their neighbors below, and their neighbors below them, etc...Each link has a metric associated with it, and the tree is pruned so that the shortest path to each destination is all that remains in the tree. Then the routing table is built based on this tree for all destinations in the network.

There are also enhancements in the basic algorithm so that not all parts of the network are told about things that are "uninteresting." This is the concept of an area, where information inside a region of routers is summarized at the border of that area and another area (usually the backbone of the network) so that the details of what's inside the area do not have to be propagated around the entire system. A protocol where there is a backbone level and areas that attach to it is called a two-level protocol, because there are two levels of information present inside the system.

Medin's Map Analogy

An analogy is very useful in trying to understand the difference between the basics of the two algorithms. Say a driver wanted to travel between Los Angeles and New York. In the distance-vector case, the driver starts out from his home, and proceeds to a gas-station, where he gets directions on which highway is the best way to get to New York. He then gets told that he should get on I-15. So he goes to I-15, and then gets off, and finds another gas station, which tells him to take I-80, etc. If I-80 were closed for bad weather, he might get told to take I-15 back towards L.A., with the idea of taking I-70. How do these gas station attendants know what the best route is? Every half hour or so, they tell their buddies about how far it is from them to various destinations around the country.

But if the gas station attendant hadn't heard I-80 was closed, he would tell our simpleminded driver to take I-15 to I-80 again. This is a routing loop! No one gas station knows anything more than the next highway to take. They know nothing of road closures farther down the route, only what their brother gas stations are telling them over the CB radio. If some trucker is busy talking on the CB, or interference drowns out a message, it's no big deal, because he'll just send the same information again the next update period. If a gas station closes because he runs out of unleaded, sooner or later his buddies will assume that he's no longer knowledgeable about what's going on, and not send drivers down the highway to him.

In the link-state case, the situation is a little different. In this case our driver again is going from Los Angeles to New York, but instead of asking each gas station attendant what the next interchange is, he has a map of the whole highway system, and looks at the road distances and calculates the shortest path through the system. If he hears on the radio that I-80 is closed, he marks that on the map, and proceeds on the next shortest path.

The use of *areas* can also be illustrated in this analogy. Say our driver has rented a car for this excursion. He gets the usual rental car company maps. He has a detailed map of the L.A. area, and a national highway system map, and a map of the New York area. On the initial leg of the journey, he uses his detailed map of the L.A. area to find a way to the nearest Inter-state highway. Once he gets to this highway, he uses the national highway system map to get into the general vicinity of New York. When he gets there, he breaks out his detailed map of the New York area to find his friend's house. In this analogy, the individual detailed maps of New York and L.A. correspond to areas, and the national highway system map corresponds to the backbone. Road closures inside Manhattan are not announced outside of New York, but are important when you get near New York.

Also note that if all the driver has when he leaves his house in L.A. is the detailed L.A. area map, he has no idea what Interstate he should get on. If he has the national map too, he can optimize his choice of highway to leave the L.A. area to get to New York. Otherwise, he would just get on the closest highway that leaves town and wait to pick up a national map at a gas station on the edge of town. This is the case when no information from outside the area is propagated inside an area. You simply head for the fastest way out, and when you get there, you figure the best route to the next area, and again, get to the border of the area, and get more detailed information.

OSPF

OSPF is a link-state protocol. Its a two-level hierarchy consisting of a backbone and multiple attached areas. When an OSPF router begins to communicate to another OSPF router, the routers "synchronize" their topology databases to make sure they have the same view of the network topology before forwarding packets through the system. If the topology is incomplete, then routing may be suboptimal or loops may occur. So it's vital that this synchronization be very robust. Likewise, during the operation of the protocol, as new links are being added and taken away, or routers crash or are rebooted, the updates that carry news of these topology changes must be flooded through the whole network, and reliably flooded to ensure synchronization of all the routers' views of the network topology. This must work even in very infrequent cases, such as when a router comes up, begins advertising a route, crashes, reboots, synchronizes, and begins readvertising the route before the first advertisement has been flooded through the system! This is one reason why link-state protocols tend to be complicated and difficult to implement in a robust way.

OSPF is very complex. Its specification is over a hundred pages of postscript output (the pictures and illustrations make it nearly impossible to read in plain text), and the average implementation is running about 9000 lines of source code (in **C** of course). Clearly, a detailed discussion of the inner workings of the protocol could be the subject of a book in itself, but let's try and discuss some of the more important aspects of the protocol and how they can be used in the real world.

continued on next page

OSPF *(continued)*

Areas

In many cases, there are concentrations of detailed information that really do not need to be passed around the whole system. Areas can be used to limit the spread of this information, and only send what's really required. For example, a corporate network may be national in scope, with a backbone consisting of a number of routers with long-haul leased lines connecting up a set of campuses which have a number of buildings with complex networks inside of each. One could take each campus, subnet a class B network inside of each campus, and run that as an OSPF area. This way information about what's inside the campus stays inside, and only the summary information (a route to the Class B network itself) escapes outside. If all the backbone links come into one building, then one can set up the area as a "stub" area, where only a default route from the backbone router is sent into the area, and all the information about the backbone itself is kept outside. OSPF further allows this summarization to occur not just to the natural mask of a subnetted network, but also to any mask. This allows a large corporation with a class A or B net to allocate "clusters" of subnets at campuses, and only advertise a single route from the cluster to the backbone.

Performance

Performance in a routing protocol means many things. Certainly one measure is the amount of routing overhead the protocol levies on the network links to distribute information around in a timely way. Because OSPF is a link-state protocol, it is very efficient about using bandwidth. Once an adjacency has been established between two routers, and the routers are synchronized, only hello messages are sent, along with routing changes as they occur. In a relatively stable network, this means very little traffic indeed. Simulations have shown that if 5% network overhead is acceptable, one can support 2000 networks worth of routing information over 9600bps links. Of course, things get busier if links are coming and going, but in the normal case, OSPF is downright miserly about link bandwidth consumption. Additionally, areas can be used to further reduce OSPF routing overhead.

Another metric of performance is how quickly the system converges back to stable routing after a link fails. Real world experience in relatively complex networks shows that OSPF can reroute about line failures in around 3 seconds. This level of performance is considered a success by most network engineers. All link-state protocols should by their nature converge quite quickly, and OSPF is not an exception.

Variable length masks

OSPF allows variable length masks and non-connected subnets to be used effectively. In OSPF, all routes are passed around with a mask associated with the route. There is no dependency on Class A,B, or C nets inside the protocol. OSPF allows a campus to use a large subnet (like a 4 bit subnet mask of a Class B net) on a large bridged campus backbone with many hosts on it, and smaller (e.g., 8 bit mask) subnets attached to the backbone contain smaller workgroups. This can really pay off if one is performing transitions from a bridged to routed environment without all the ugly kludges of running multiple logical subnets on a single physical network. OSPF also doesn't mind subnets of a network being separated by pieces of another network. This means that a regional network could transit a local campus network between two of the regional's routers, if that was desired, without coming up with a dummy subnet to run in parallel with the campus addressing on the intervening network.

It should be pointed out that to make full use of this feature, you need more than just OSPF. The IP forwarder code of the router in question must be able to route on best-match, or some other scheme that supports variable length subnetting. The protocol assumes the forwarder can deal with this, so if a router doesn't handle this for some reason, then one had better be careful in how the OSPF system is configured to make sure this feature isn't used! Fortunately, almost all routers which implement OSPF support this functionality, and it is fairly clear that this will be required of routers used in most environments in the future, as people are even modernizing older protocols such as RIP to support this feature.

Efficient LAN behavior

In a pure link-state protocol, each router advertises a link to each router it can communicate with. On the point to point lines, this works fine, but on LAN's like ethernet or FDDI systems, where one can find large numbers of routers that can all directly communicate with each other, there would be a lot of links being advertised. So to reduce the amount of information that must be passed around, OSPF has something called a *Designated Router* or DR for short. Instead of all the routers sending information about themselves to each other, a designated router is elected (based on a priority setting), and all the other routers communicate with the DR, and the DR then redistributes the information from the other routers to each other. If the DR should crash, a new DR has to be elected, synchronized with, etc... Since people tend to be building networks out of large FDDI type backbones with large numbers of routers attached, the DR needs to be very reliable to ensure the network can continue to talk properly. This is why OSPF also has something called a "Backup Designated Router" or Backup DR. The Backup DR communicates with all the other routers on the LAN, just like the DR does, but if the DR should fail, the routers can very quickly switch to the Backup DR to ensure the network stays up and operational. While Dual IS–IS does support the use of a DR on LANs, it does not support the concept of a Backup DR.

Additionally, OSPF only uses the router priority values when a new DR or Backup DR has to be elected. If a new router becomes adjacent to the DR and has a higher (more preferable) router priority, then no election process is initiated. Only when the existing DR fails does the new router with the higher priority get elected as the DR. Dual IS–IS will elect a new DR any time a router with a higher priority tries to become adjacent. This can lead to an unnecessary amount of DR transitions occurring, which decreases the stability and robustness of the network. It can also lead to thrashing if the new DR runs into a memory, or other problem, which causes it to crash when becoming a DR, but works fine otherwise.

OSPF also uses IP multicasting to send updates on LANs. The multicasts also have the *Time-To-Live* (TTL) field set to 1, to ensure that if some miscreant router or host would try and forward an OSPF packet, that the TTL would be decremented to 0 and an ICMP error message would be sent back. This has actually happened in practice, and just goes to show that robustness and fault tolerance really needs to be designed in any protocol for use in the Internet. Hosts on the network never "see" the OSPF messages (because they are multicast, not broadcast), and poorly configured hosts don't have the option of responding in funny ways.

Efficient support of non-broadcast networks

OSPF was designed to deal with non-broadcast multi-access networks like X.25 PDNs, Frame Relay, the Defense Data Network (DDN), Hyperchannel A-series adapters, etc...in a robust and reliable way.

continued on next page

OSPF *(continued)*

Most routing protocols handle only two basic types of nets: broadcast and non-broadcast. On broadcast nets, updates are multicast on the network. On non-broadcast nets, routers all unicast messages to each other just like on a point-to-point link. Non-broadcast multi-access nets are basically treated as point-to-point lines. In order to avoid extra hops, each router must be configured with a point-to-point style link to all other routers in the system. Thus the same piece of routing information may traverse the network many times between the same routers. This can be very uneconomical when billing by packet usage is occurring.

OSPF handles these types of networks by modeling them as LANs, with DRs and backup DRs. But since it can't elect the DR by multicasting hello messages, each router maintains a small list of manually configured potential DRs. These routers are then polled when a new router comes up and a DR found or a new one elected. This can mean much more efficient behavior. Of course, the routers must be able to talk to each other, and if they can't, then you can get yourself into trouble. If network failures of this type are commonplace, then OSPF can run over these types of nets in a more conventional point-to-point mode.

Security

OSPF is the first Internet routing protocol with authentication designed in from the start. All OSPF packets carry a 64 bit authentication field in their headers, and this field is checked first before anything is done with the packet. Routers can set different types of authentication in different areas, and different keys can be used on each link. Routers operating on LANs can use this feature to prevent unauthorized routers from communicating with the OSPF system. Serial lines in OSPF do not need to be numbered, so proper use of keys can prevent accidental misconfigurations of serial lines. And multiple independent OSPF systems can co-exist on multi-access networks by using different keying or authentication schemes.

External network support

OSPF was designed for use in the global Internet. It has many features which are very useful when dealing with connections to other ASs. One of these features is the ability to "tag" information about routes that are being imported into the system. OSPF considers all routes learned by other means external (including static routes). Each of these external routes carries a tag with it that is usually filled with the AS number of the external system delivering the routes to you. This allows you to make AS border routing policy based on the AS level and not purely on a net by net basis. In complex transit systems, this feature is a life saver!

OSPF also supports multiple ways of dealing with external routes. The external routes can be imported with a metric (manufactured at the border gateway) which is defined as being comparable to the internal OSPF metric (type 1) or a metric which is always defined to be greater than any internal OSPF metric (type 2). For example, if a network always wishes to prefer one path to an external network over a less direct backup path, type 2 routes can be used to always route to the primary exit. If one wants to minimize the internal cost and find a way out to the closest exit gateway, then type 1 routes can be selected. This may make sense if a regional network has 2 interconnects to the NSFnet backbone, and wishes to pick one exit for one half of the system, and the other for the second half to equalize the load.

OSPF also has a built-in trust model to support "protection" of routing information. Intra-area routes are always preferred over inter-area routes, which are always preferred over external type 1 routes which are preferred over external type 2 routes. Thus, internal campus routes can never be overridden from the backbone, and one AS's internal routes can never be overridden from the outside. However, OSPF has very few firewalls internal to the AS itself. It really is a IGP in the strictest sense, and all route filtering and firewalling really has to occur on the edges of the OSPF system and not within it.

Type of Service (TOS) support

OSPF fully supports IP TOS options, i.e., the three TOS bits in the IP header. OSPF can be set up with multiple metrics, one for each TOS, to match network topology to application class of service. It's thus possible to route bulk USENET traffic over cheap high delay satellite links, and interactive traffic over more expensive low delay terrestrial links. While it's true that few hosts support this now, it is hoped that with a routing protocol that can finally support TOS properly being standardized and deployed, hosts will have an incentive to properly set the TOS bits when building packets.

Comparisons

It is often useful to compare the operation of a one protocol against others that are in use to judge whether or not the protocol design has been successful or not. In the Internet today, there are a large number of IGPs in use. Most of them, such as RIP and Cisco System's IGRP, are based on distance vector algorithms. There are also groups working on another link-state IGP which combines ISO *Connectionless Network Service* (CLNS) support with IP support in one routing protocol, called Dual IS–IS. Because OSPF and Dual IS–IS are both link-state protocols, they have many things in common.

Performance

Because RIP and IGRP are distance-vector protocols, periodically, they transmit updates to their neighbors, basically full dumps of their routing tables. This occurs every 30 seconds for RIP and, by default, 90 seconds for IGRP. If this time is increased, the time it takes for the system as a whole to converge is significantly increased. If this time is decreased, then the protocol's overhead is increased. Since OSPF only sends out incremental updates as links' state changes, and a periodic refresh every 30 minutes, its overhead is very low. Statistics from operational deployments in the NASA and BARRNet systems indicate significantly lower overhead compared to RIP, which was noticed after transitioning from RIP to OSPF. When IGRP systems begin to transition to OSPF, great savings are also expected to occur, based on current experience with the two protocols passing equivalent numbers of routes. Convergence times for OSPF based systems are also very quick, on order of 3 seconds or so in large systems. RIP and IGRP systems can take considerably longer to converge, on order of minutes in many cases, causing disruptions in network service to users.

Variable length mask support

Neither RIP nor IGRP have the ability to support variable length masks effectively because neither protocol passes mask information with the route itself. This means the mask must be derived from other information in the router, usually an interface mask or some other state information. It also makes it impossible to support disconnected subnets. Use of these features in the operational Internet have found very valuable and useful, especially as more and more bridged nets become routed.

LAN behavior

All the major IGPs are relatively efficient about LAN bandwidth use. OSPF, however, additionally provides detection and isolation of "diode" situations where one station can hear another but not successfully transmit data to other stations.

continued on next page

OSPF *(continued)*

This can cause black holes in routing in RIP and IGRP systems when diode failures occur. Also, RIP and IGRP broadcast on LANs, resulting in hosts getting the packets and causing interrupts in processing and possible garbage replies and broadcast storms in poorly configured environments. Since OSPF multicasts all updates, this is not a problem.

Non-broadcast network support

RIP, IGRP, and also the Dual IS–IS protocol have to support non-broadcast multi-access networks basically as point-to-point links. OSPF offers the option of using a much more efficient way of supporting such networks, resulting in significant monetary savings in many cases. There are some workarounds that can be performed, but they usually end up with extra hops being introduced.

Security

Neither RIP or IGRP support authentication as part of the protocol. While some controls exist for ignoring updates from selected routers, no effective solutions exist without internal protocol support. OSPF currently has 2 standard authentication methods supported, none and simple password, with additional methods being defined in future appendices, and support for private non-standard authentication methods. Dual IS–IS has recently been modified to support a form of authentication, however the authentication information is basically carried as an option, and is not built into the header of each routing packet, as in the case of OSPF. This can result in a considerably weaker authentication capability than what is possible in OSPF.

External network support

RIP has no tagging support at all. Multiple parallel IGRP sessions can accomplish some effects of tagging, but very few of these sessions can effectively be run on a given router, making extensive use of this feature relatively expensive. Neither protocol supports a hierarchical trust model similar to what OSPF uses. The original Dual IS–IS specification and the base IS–IS protocol had very limited capabilities of dealing with external routes, but have added some capabilities, such as the type 1 and type 2 external route features of OSPF, and is currently being modified to support next hop optimization and other capabilities found in OSPF V2.

Area support

OSPF and IS–IS are very similar in many ways. OSPF, however, allows external routing information to be propagated into areas, or not (in which case they are OSPF stub areas). This allows a network manager the option of trading off memory inside of routers in an area against the advantages of picking a more optimal exit for packets leaving the area. Note that if simplifications are made in determining the exit router, it is possible for asymmetric paths to be taken unintentionally inside an AS. Dual IS–IS does not allow this choice; all IS–IS areas are effectively stub areas. The designers of OSPF believed that there would be many topological configurations in which the exit router optimization would be a significant advantage, and many real world network operators appreciate and use this option.

Virtual Links

Both Dual IS–IS and OSPF operate with a two level hierarchy of routing information. There is a backbone area, and other areas that attach to the backbone. Both protocols critically depend on the backbone area not partitioning to ensure reliable routing inside the AS. In many topologies, there may be links that can be used to heal a partition in the backbone area that cannot be used because they are part of non-backbone area. OSPF has the concept of a Virtual Link, which is a virtual backbone link that can be used to keep the backbone connected from a routing point of view, where the area configuration would not normally have supported this.

Dual IS–IS does not have this feature, and thus when trying to ensure a routing configuration that will keep the backbone connected in the face of link failures, the network manager would have to settle for a less than optimal area configuration, or possibly as less robust network configuration than would have been possible with OSPF.

Conclusions

OSPF has been recommended by the IETF to advance to a Draft Standard, which normally progresses to a Full Standard in a relatively short period of time. The IETF Router Requirements work group has decided to have all router vendors that support a dynamic routing protocol be required to support OSPF to ensure router interoperability with relatively high functionality. This will mark the first time an IGP will be required for routers to support, and should greatly decrease the hassle of running a multi-vendor network. The University of Maryland has built a network simulation system to model OSPF's behavior in large systems, and this along with operational experience have given us reason to believe that OSPF will scale nicely into large systems.

Several OSPF interoperability tests have been conducted so far, and more are planned. In the most recent one, held at FTP Software, something like a dozen implementations interoperated with each other, and every major vendor of IP routers participated. Even DEC, who is the chief Dual IS–IS proponent, has stated that they will support OSPF. It has become the routing protocol of choice for IP networks, and is being used in a production mode in several large backbone, regional, and corporate networks internationally. It is an example of how network engineers and architects have developed a protocol from a requirements definition phase, to an architecture phase, to a design phase, to an implementation and test phase, and finally through interoperability testing and wide scale deployment that is an example for other protocol development efforts to follow. OSPF has truly become the protocol of choice for high performance routing support.

References

[1] Moy, J., "The OSPF Specification," RFC 1131, October 1989.

[2] Moy, J., "The OSPF Specification, Version 2," Internet Draft, January 1991.

[3] Corporation for National Research Initiatives, "Proceedings of the Seventeenth Internet Engineering Task Force," Pittsburgh Supercomputing Center, May 1–4 1990.

[4] Corporation for National Research Initiatives, "Proceedings of the Twentieth Internet Engineering Task Force," Washington University, March 4–8 1991.

[5] Hedrick, C., "The Routing Information Protocol," RFC 1058, June 1988.

[6] Moy, J., "The OSPF V2 Specification," RFC 1247, August 1991.

[7] Callon, R., "Use of OSI IS–IS for Routing in TCP/IP and Dual Environments," RFC 1195, December 1990.

[8] Dern, D., "Standards for Interior Routing Protocols: They're emerging, but still under debate and development," *ConneXions,* Volume 4, No. 7, July 1990.

MILO MEDIN is Network Architect for the NASA Science Internet Office at NASA Ames Research Center, and heads up engineering for the NASA component of the NREN. He is an active member of the Internet Engineering Task Force and a member of the working group which designed OSPF. Although not explicitly stated in this article, Milo is a strong advocate of Ships In the Night routing.

The Deployment of Privacy Enhanced Mail

by James M. Galvin, Trusted Information Systems

Introduction

The *Privacy Enhanced Mail* (PEM) specifications are the culmination of several years work by the Privacy and Security Research Group of the Internet Research Task Force. PEM provides for the addition of three, and sometimes four, security services: message origin authentication, message integrity, message confidentiality, and, when asymmetric key management techniques are employed, non-repudiation. The services are provided end-to-end by RFC 822 conformant user agents; no changes are required of intermediate, relay mail systems.

TLCA

During the recent Atlanta *Internet Engineering Task Force* (IETF) meeting a new PEM working group was created to oversee the progression of the current revisions to the specifications through the IETF standards process. The published specifications are a series of three RFCs: message processing, [1] certificate-based key management, [2] and algorithm identifiers. [3] The revisions will be a series of four RFCs, [4, 5, 6, 7] with a few ancillary RFCs referenced. The new, fourth specification, defines the services provided by and the communication interface to a *Top Level Certification Authority* (TLCA).

In addition, a reference implementation of the specifications has been under development for the past two years, jointly by Trusted Information Systems, Bolt Beranek and Newman, and RSA Data Security Incorporated. There are approximately 20 beta test sites with all but a few running Version 5.0 Beta. At INTEROP 91, there will be a demonstration of the reference implementation. Pending final revisions, the reference implementation will be made openly available to the Internet community.

The questions that remain to be answered for the Internet community are, when will the reference implementation of PEM be deployed, and how will it be deployed. The deployment of the reference implementation of PEM involves more than the development of a software system that can be made openly available via anonymous FTP. There are technical as well as policy issues that must be resolved, some of which are outside the scope of the specifications. Some of these are discussed below, in the schedule of events leading to deployment. All issues and status reports are posted and discussed on the PEM developers electronic mailing list; to join pem-dev@tis.com send a message to pem-dev-request@tis.com.

Schedule

The set of events culminating in the deployment of the reference implementation of PEM is as follows:

- Complete the software development, except for the remaining unresolved technical issues (see next bullet). This should be done in time for the INTEROP 91 demonstration.

- Resolve the outstanding technical issues. The principal technical issue remaining is the uniform trust requirement. The current revision to RFC 1114 specifies a requirement for a uniform level trust throughout the certificate-based key management infrastructure. There was a lively discussion of this issue by the members of the IETF PEM working group, but unfortunately no consensus was reached. One mechanism by which it is proposed to enforce the uniform level of trust is to restrict the distinguished name forms allowed in a certificate used by PEM. The PEM software would be required to validate the name forms. This has the advantage that it will not be necessary for users to understand certification hierarchies and the level of trust that may or may not be associated with them.

It has the disadvantage that the required trust level may be more than some organizations practice during their normal course of business.

Another mechanism by which it could be enforced is to require a contractual agreement to be executed between organizations issuing certificates and their respective TLCAs. This has the advantage of providing a forum in which non-conforming organizations may be dealt with fairly. It has the disadvantage that smaller organizations may not be prepared, e.g., for fiscal reasons, to execute the contract. Interim meetings and discussions on the PEM developers mailing list are expected to explore this issue sufficiently so that consensus should be possible at the Santa Fe IETF meeting.

- Progress the revised specifications onto the IETF standards track as Proposed Draft Standards in order to encourage other implementations. This should happen at the Santa Fe IETF meeting. If necessary, consider progressing only the stable specifications if consensus cannot be reached on the outstanding technical issue.

- Resolve the outstanding policy issue. The principal policy issue remaining is the export restriction of the software. PEM uses *cryptography* to support the message origin authentication and message confidentiality services, which have export restrictions associated with them. This may make anonymous FTP an inappropriate mechanism with which to distribute PEM. One possible resolution is to continue the necessary formal process and make PEM available only within the U.S. initially. Implementations outside of the U.S. are known to be in progress, but their availability is uncertain.

- Determine a mechanism suitable for making PEM available as easily as possible to as broad a community of users as possible. A suitable mechanism with which to deploy PEM that addresses the export restrictions is still under discussion.

References

[1] Linn, John, "Privacy Enhancement for Internet Electronic Mail: Part I—Message Encipherment and Authentication Procedures," RFC 1113.

[2] Kent, Steve & John Linn, "Privacy Enhancement for Internet Electronic Mail: Part II—Certificate-Based Key Management," RFC 1114.

[3] Linn, John, "Privacy Enhancement for Internet Electronic Mail: Part III—Algorithms, Modes, and Identifiers," RFC 1115.

[4] Linn, John, "Privacy Enhancement for Internet Electronic Mail: Part I—Message Encipherment and Authentication Procedures," Internet Draft: `draft-ietf-pem-msgproc-00.txt`.

[5] Kent, Steve, "Privacy Enhancement for Internet Electronic Mail: Part II—Certificate-Based Key Management," Internet Draft: `draft-ietf-pem-keymgmt-00.txt`.

[6] Balenson, David, "Privacy Enhancement for Internet Electronic Mail: Part III—Algorithms, Modes, and Identifiers," Internet Draft: `draft-ietf-pem-algorithms-00.txt`.

[7] Kaliski, Burton S, "Privacy Enhancement for Internet Electronic Mail: Part IV—Notary, Co-Issuer, CRL-Storing and CRL-Retrieving Services," Internet Draft: `draft-ietf-pem-notary-00.txt`.

[8] Dern, D., "Interview with Steve Kent on Internet Security," *ConneXions,* Volume 4, No. 2, February 1990.

[9] Galvin, J. "Components of OSI: The Security Architecture," *ConneXions,* Volume 4, No. 8, August 1990.

JAMES M. GALVIN is a Senior COMSEC Scientist at Trusted Information Systems (TIS). Dr. Galvin's responsibilities emphasize communications security, especially computer networks, architectures, policies, and procedures. He is a principal in the development of TIS' soon to be openly available implementation of Privacy Enhanced Mail. He is very active in the IETF Security Area Advisory Group and Chair of the OSI Implementor's Workshop Security Special Interest Group, hosted quarterly by the National Institute of Standards and Technology. He received his Ph.D. and M.S. degrees, both in Computer Science, from the University of Delaware in 1988 and 1986, respectively. In 1982, he received his B.S. in Computer Science and Mathematics from Moravian College in Bethlehem, PA.

CONNEXIONS
480 San Antonio Road
Suite 100
Mountain View, CA 94040
415-941-3399
FAX: 415-949-1779

ADDRESS CORRECTION
REQUESTED

FIRST CLASS MAIL
U.S. POSTAGE
PAID
SAN JOSE, CA
PERMIT NO. 1

CONNEXIONS

EDITOR and PUBLISHER Ole J. Jacobsen

EDITORIAL ADVISORY BOARD

Dr. Vinton G. Cerf, Vice President,
Corporation for National Research Initiatives

A. Lyman Chapin, Chief Network Architect,
BBN Communications

Dr. David D. Clark, Senior Research Scientist,
Massachusetts Institute of Technology

Dr. David L. Mills, Professor,
University of Delaware

Dr. Jonathan B. Postel, Communications Division Director,
University of Southern California, Information Sciences Institute

Subscribe to CONNEXIONS

U.S./Canada ❑ $150. for 12 issues/year ❑ $270. for 24 issues/two years ❑ $360. for 36 issues/three years

International $ 50. additional **per year** **(Please apply to all of the above.)**

Name ______________________ Title ______________________

Company ______________________

Address ______________________

City ______________________ State ____________ Zip ____________

Country ______________________ Telephone (   ) ______________

❑ Check enclosed (in U.S. dollars made payable to CONNEXIONS).
❑ Visa ❑ MasterCard ❑ American Express ❑ Diners Club Card # ______________ Exp.Date ________

Signature ______________________

Please return this application with payment to:

CONNEXIONS
480 San Antonio Road, Suite 100
Mountain View, CA 94040 U.S.A.
415-941-3399 FAX: 415-949-1779

Back issues available upon request $15./each
Volume discounts available upon request